Dr. Joseph Teguh Santoso, S.Kom, M.Kom

CARA MEMANIPULASI PEMBELAJARAN MESIN

(Machine Learning)

YAYASAN PRIMA AGUS TEKNIK

CARA MEMANIPULASI PEMBELAJARAN MESIN

(Machine Learning)

Dr. Joseph Teguh Santoso, S.Kom, M.Kom

YAYASAN PRIMA AGUS TEKNIK

PENERBIT :
YAYASAN PRIMA AGUS TEKNIK
Jl. Majapahit No. 605 Semarang
Telp. (024) 6723456. Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

ISBN 978-623-8642-00-7 (PDF)

CARA MEMANIPULASI PEMBELAJARAN MESIN (Machine Learning)

Penulis :
Dr. Joseph Teguh Santoso, S.Kom., M.Kom

ISBN : 978-623-8642-00-7

Editor :
Muhammad Sholikan, M.Kom

Penyunting :
Dr. Mars Caroline Wibowo. S.T., M.Mm.Tech

Desain Sampul dan Tata Letak :
Irdha Yunianto, S.Ds., M.Kom

Penebit :
Yayasan Prima Agus Teknik Bekerja sama dengan
Universitas Sains & Teknologi Komputer (Universitas STEKOM)

Anggota IKAPI No: 279 / ALB / JTE / 2023

Redaksi :
Jl. Majapahit no 605 Semarang
Telp. (024) 6723456
Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

Distributor Tunggal :
Universitas STEKOM
Jl. Majapahit no 605 Semarang
Telp. (024) 6723456
Fax. 024-6710144
Email : info@stekom.ac.id

KATA PENGANTAR

Puji Syukur kami panjatkan kehadirat Tuhan Yang Maha Esa atas berkah-Nya yang melimpah, yang telah memungkinkan penyelesaian penulisan buku yang berjudul "*Memanipulasi Pembelajaran Mesin (Machine Learning)".* Kesenjangan antara kecerdasan mesin dan persepsi manusia masih besar, meski pembelajaran mendalam berkembang. Keamanan pembelajaran mendalam menjadi tantangan besar karena rentannya terhadap serangan musuh. Algoritma permusuhan dirancang untuk mengeksploitasi kerentanan ini dengan skenario serangan yang disimulasikan oleh musuh cerdas. Teknologi serangan pembelajaran mesin telah berkembang dalam berbagai bidang seperti visi komputer, pemrosesan bahasa alami, dan keamanan siber.

Dalam pembelajaran diskriminatif, masalah manipulasi dirumuskan dengan jaringan saraf yang mengukur perbedaan statistik antara fitur data pelatihan dan fitur data lawan. Jaringan dapat mencari ruang laten pada data pelatihan untuk membuat contoh permusuhan. Algoritma manipulasi pembelajaran mesin bervariasi dalam asumsi desain terkait pengetahuan musuh, strategi serangan, dan dampaknya. Buku ini menyajikan tinjauan literatur tentang hubungan antara adversarial learning dan serangan siber, membandingkan ancaman permusuhan dalam model pembelajaran mesin dengan vektor serangan dalam model pembelajaran mendalam. Representasi data non-stasioner dan kelas konsep yang dipelajari oleh jaringan pembelajaran mendalam yang bermusuhan juga disurvei. Taksonomi contoh serangan digunakan untuk menganalisis ketahanan sistem pembelajaran dalam lanskap yang terus berubah. Profil pembelajaran teori permainan menganalisis ketahanan sistem pembelajaran terhadap optimalisasi dinamis dan stabilitas solusinya.

Dari sudut pandang privasi data, penulis mengkaji risiko, ancaman, dan kerentanan keamanan siber terkait pelestarian privasi dan serangan fisik. Opsi deteksi dan respons disediakan untuk algoritme pembelajaran mendalam, serangan dalam sistem kompleks, pembelajaran mendalam permusuhan, pengoptimalan yang kuat, dan kontrol cerdas. Tema penelitian tersebut dapat diterapkan pada desain sistem yang tangguh dengan pengumpulan data yang menjaga privasi untuk menganalisis ancaman, metadata, dan pola serangan. Ini juga dapat digunakan untuk studi kualitas data dan asal informasi dalam penambangan data permusuhan untuk menghasilkan sistem IoT dengan pertahanan dalam algoritma pembelajaran yang menggabungkan keamanan dengan privasi. Orkestrasi keamanan kemudian dapat menyediakan solusi keamanan siber sebagai layanan di Internet untuk akses yang andal ke sistem pembelajaran mesin di dunia nyata.

Buku ini juga berfungsi sebagai referensi bagi peneliti yang ingin membandingkan skenario serangan dan mekanisme pertahanan dalam pembelajaran mendalam permusuhan. Teknik-teknik invasif yang dikenal dan tindakan penanggulangannya juga dibahas untuk meningkatkan keamanan siber di masa depan. Buku ini menyoroti bahwa masalah keamanan dalam pembelajaran mendalam terutama terletak pada formulasi matematis dan mekanisme pembelajarannya. Penulis menggunakan teori permainan untuk mengatasi sampel manipulasi

yang memengaruhi manipulasi data. Pertanyaan-pertanyaan dalam buku ini membantu dalam merancang strategi pembelajaran mendalam adversarial seputar tindakan musuh siber dalam serangan siber. Ini membantu dalam mengevaluasi dan memilih skenario ancaman siber yang berisiko tinggi dengan fokus pada deteksi dan pencegahan keamanan.

Buku ini dimulai dengan ulasan pembelajaran mesin permusuhan di Bab. 1 beserta perbandingan pendekatan baru versus pendekatan yang sudah ada terhadap pembelajaran mesin permusuhan teoritis permainan. Dalam bab 2 memposisikan kontribusi buku ini kami berbeda dengan literatur terkait tentang (i) mekanisme keamanan permusuhan dan jaringan permusuhan generatif, (ii) contoh permusuhan untuk pengklasifikasi mendalam yang menyesatkan dan model pembelajaran mendalam permusuhan teoritis permainan, dan (iii) contoh permusuhan dalam pembelajaran transfer dan adaptasi domain untuk keamanan siber. Serangan permusuhan yang muncul karena pengorbanan keamanan dan privasi dalam pembelajaran mendalam diberikan dalam Bab. 3. Mereka merangkum permukaan serangan siber, fisik, aktif, dan pasif dalam lingkungan kritis keamanan yang saling bergantung, saling terhubung, dan interaktif untuk sistem pembelajaran. Permukaan serangan seperti itu meningkat secara vertikal dalam jumlah, volume, dan secara horizontal dalam jenis, fungsionalitas melalui Internet, jejaring sosial, ponsel pintar, dan perangkat IoT. Keamanan otonom dalam strategi mitigasi ancaman yang melindungi diri dan menyembuhkan diri sendiri harus mempertimbangkan permukaan serangan tersebut dalam mekanisme kontrol domain jaringan untuk mengidentifikasi ancaman dan memilih metode pembelajaran mesin dan penambangan data yang sesuai untuk pembelajaran permusuhan.

Bab 4 membahas pembelajaran mendalam permusuhan dalam teori permainan, membandingkan algoritma komputasi dengan optimasi stokastik. Contoh-contoh digunakan untuk membangun musuh yang sensitif terhadap biaya penambangan data. Kuantifikasi hipotesis membawa kita pada masalah fungsional, orakular, pengambilan sampel, dan optimasi dalam permainan pembelajaran. Teori kompleksitas sampel dan automata fuzzy digunakan dalam model manipulasi. Dinamika pengambilan sampel ini dapat diterapkan dalam pemrosesan sinyal untuk keamanan siber.

Bab 5 menyajikan teori dan algoritma untuk pembelajaran mendalam permusuhan. Algoritme ini juga dapat digunakan untuk memeriksa konsistensi dan penerapan spesifikasi sistem pembelajaran untuk menggabungkan data serangan dan memperkuat spesifikasi menjadi model pembelajaran permusuhan baru dengan metrik penilaian kerentanan, protokol, dan fusi penanggulangan. Contoh penerapan serangan manipulasi karena pembelajaran mendalam teoretis permainan yang diusulkan dalam buku ini disajikan di Bab 6. penulis berfokus pada spam statistik dan sistem otonom dengan gambar dan video, namun menemukan literatur yang relevan dalam berbagai aplikasi keamanan siber, seperti kriptanalisis, steganalisis, dan deteksi phishing. Meskipun teori permainan dapat digunakan untuk pemodelan generatif-diskriminatif dalam manipulasi pembelajaran mendalam, implementasinya seringkali sulit dalam arsitektur dangkal untuk pembelajaran mesin.

Bab 7 membahas pemanfaatan pembelajaran adversarial dalam teknologi privasi. Penulis mengusulkan definisi ontologi untuk kepercayaan, ketahanan, dan kelincahan

terhadap agen ancaman. Mereka menekankan perlunya teknik penambangan data yang menjaga privasi untuk melengkapi manipulasi pembelajaran mendalam. Ontologi metrik keamanan dan ketergantungan diusulkan untuk memodelkan agen AI bermusuhan dalam strategi serangan siber. Dengan menggunakan teori permainan, penulis mencoba mengatasi kesulitan komputasi dalam mengukur utilitas dan kehilangan informasi terkait. Mereka menegaskan bahwa agen AI bermusuhan dapat meningkatkan respons terhadap ancaman keamanan siber melalui pemindai cerdas, firewall, anti-malware, dan teknologi lainnya

Demikian buku ajar ini kami buat, dengan harapan agar pembaca dapat memahami informasi dan juga mendapatkan wawasan yang komprehesif mengenai , praktik manipulasi dalam pembelajaran mesin. Semoga buku ini memberikan kontribusi yang berarti bagi pembaca dalam memahami dan mengatasi tantangan yang kompleks dalam dunia pembelajaran mesin. Terima Kasih.

Semarang, Mei 2024
Penulis

Dr. Joseph Teguh Santoso, S.Kom., M. Kom.

DAFTAR ISI

BAB 1
PEMBELAJARAN MESIN

Bab ini menyelidiki kesenjangan kekuatan antara kecerdasan mesin dan persepsi manusia dalam pembelajaran mesin untuk keamanan dunia maya dengan algoritma pembelajaran adversarial teoritis permainan. Dalam bab ini, kami akan melakukan tinjauan literatur untuk memberikan wawasan baru tentang hubungan antara manipulasi pembelajaran mesin dan keamanan siber. Kami berupaya mensurvei dan merangkum representasi data non-stasioner yang dipelajari oleh model pembelajaran mesin. Ketahanan pemodelan harus disurvei untuk menghasilkan ringkasan contoh-contoh adversarial dan algoritma-algoritma adversarial. Kami juga akan mensurvei penggunaan optimasi cembung, optimasi stokastik, dan komputasi evolusioner dalam formulasi pembelajaran mendalam yang bermusuhan. Studi menarik lainnya adalah mekanisme pertahanan yang tersedia untuk model pembelajaran mendalam yang diterapkan di lingkungan dunia nyata.

Penambangan data adalah studi tentang pembelajaran pola matematika secara otomatis dari informasi dalam database. Ini adalah proses penemuan pengetahuan yang memerlukan pengembangan algoritma komputasi untuk prapemrosesan, pemodelan, dan pascapemrosesan data yang diberikan sistem basis data. Namun, desain algoritme tersebut harus didasarkan pada paradigma pembelajaran mesin. Paradigma pembelajaran mesin adalah mode pembelajaran komputasi berdasarkan beberapa asumsi statistik yang mendasarinya, seperti tingkat pengawasan manusia dalam data pelatihan atau distribusi data yang mendasarinya. Contoh paradigmanya meliputi pembelajaran yang diawasi, pembelajaran tanpa pengawasan, pembelajaran semi-supervisi, pembelajaran penguatan, pembelajaran meta, dan pembelajaran mendalam.

Asumsi statistik standar, yang disebut asumsi stasioneritas, adalah bahwa data pelatihan yang digunakan oleh model untuk mempelajari pola matematika dan data pengujian yang digunakan untuk mengevaluasi seberapa baik model tersebut mengenali pola-pola tersebut diambil sampelnya dari distribusi probabilitas dasar yang sama yaitu independen dan terdistribusi identik. (i.i.d) variabel acak. Namun asumsi stasioneritas tidak berlaku di sebagian besar aplikasi dunia nyata; data pelatihan dan pengujian jarang memiliki distribusi yang persis sama dan tidak sering kali bersifat i.i.d. Oleh karena itu, paradigma pembelajaran yang kuat untuk analisis data non-stasioner telah menjadi tujuan manipulasi pembelajaran mesin. Pembelajaran manipulasi memiliki aplikasi di berbagai bidang seperti pemfilteran spam, deteksi virus, deteksi intrusi, deteksi penipuan, otentikasi biometrik, verifikasi protokol jaringan, periklanan komputasi, sistem pemberi rekomendasi, penambangan web media sosial, pemodelan kinerja sistem yang kompleks, dan sebagainya.

Algoritma manipulasi pembelajaran (atau nantinya kita akan sering menyebut dengan pembelajaran adversarial) mesin dirancang khusus untuk mengeksploitasi kerentanan dalam algoritma pembelajaran mesin tertentu. Kerentanan ini disimulasikan dengan melatih algoritma pembelajaran dalam berbagai skenario dan kebijakan serangan. Skenario serangan

diasumsikan dirumuskan oleh musuh yang cerdas, dan kebijakan serangan yang optimal adalah kebijakan yang dapat menyelesaikan satu atau banyak masalah optimasi pada satu atau banyak skenario serangan, dengan memperhatikan bahwa berbagai algoritma manipulasi pembelajaran mesin mungkin berbeda dalam asumsi statistiknya. atas pengetahuan musuh, pelanggaran keamanan, strategi serangan, dan pengaruh serangan.

Dengan demikian, algoritma pembelajaran yang telah dirancang untuk mengimbangi serangan menjadi kuat terhadap serangan tersebut; kerentanannya tidak lagi rentan. Dengan demikian, tujuan pembelajaran adversarial dapat dianggap sebagai salah satu menemukan solusi untuk fungsi tujuan dalam algoritma pencarian dan optimasi yang bertahan dari skenario serangan. Setelah ditemukan, solusi ini dapat dimasukkan ke dalam desain banyak algoritma pembelajaran mesin sebagai mekanisme pertahanan untuk mencegah serangan. Pembelajaran mendalam mengacu pada kelas algoritma jaringan saraf tertentu. Algoritma ini terdiri dari banyak tahapan pemrosesan informasi non-linier dalam arsitektur hierarki yang dieksploitasi untuk klasifikasi pola dan pembelajaran fitur. Penelitian pembelajaran mendalam bertujuan untuk menemukan algoritma pembelajaran mesin pada berbagai tingkat abstraksi data.

Pembelajaran mendalam dengan data berdimensi tinggi terbukti rentan terhadap serangan atau hacker. Serangan semacam itu dibuat berdasarkan pengetahuan sebelumnya, observasi, dan eksperimen terhadap fungsi kerugian dalam model pembelajaran mendalam. Investigasi sistematis terhadap desain fungsi kerugian pembelajaran mendalam untuk bertahan melawan musuh adalah bidang penelitian baru dan praktis. Selain itu, analisis kesalahan statistik pada fungsi kerugian berbasis data harus mempertimbangkan tujuan pengoptimalan yang bertentangan dalam model yang diserang, seperti akurasi, skalabilitas, waktu proses, dan keragaman, yang ditentukan berdasarkan distribusi data yang mendasarinya. Fungsi kerugian telah didefinisikan dalam konteks berbagai paradigma pembelajaran mesin yang berlaku untuk pembelajaran mendalam. Dalam pembelajaran yang diawasi, fungsi kerugian didefinisikan sebagai kriteria yang sesuai untuk estimasi probabilitas kelas. Dalam pembelajaran statistik, fungsi kerugian didefinisikan sebagai minimalisasi risiko empiris dalam data pelatihan. Dalam pembelajaran komputasi, fungsi kerugian dikatakan meminimalkan aturan keputusan Bayes untuk prediktor dengan menghitung kemungkinan kesalahan klasifikasi yang diharapkan. Model pembelajaran berbasis energi adalah kerangka teoritis untuk inferensi statistik dan pembelajaran komputasi yang bercirikan fungsi kerugian.

Masalah pembelajaran adversarial dalam fungsi kerugian pembelajaran diskriminatif biasanya dirumuskan dengan metrik perbedaan statistik antara fitur data pelatihan dan fitur data adversarial. Ruang laten pada data pelatihan berdimensi tinggi juga dapat dicari oleh jaringan dalam untuk membangun contoh adversarial. Bergantung pada tujuan, pengetahuan, dan kemampuan musuh, contoh adversarial juga dapat dibuat berdasarkan pengetahuan sebelumnya, observasi, dan eksperimen terhadap fungsi kerugian dalam pembelajaran mendalam. Oleh karena itu, algoritma pembelajaran adversarial yang ada berbeda dalam asumsi desain mengenai pengetahuan musuh, strategi serangan, pengaruh serangan, dan pelanggaran keamanan. Selain itu, contoh-contoh serangan diketahui

berpindah antar keragaman model pembelajaran mesin yang spesifik data. Oleh karena itu, kinerja prediktif model pembelajaran mendalam yang sedang diserang merupakan bidang yang menarik untuk diteliti.

Teknologi serangan adversarial ada dalam visi komputer, pemrosesan bahasa alami, keamanan dunia maya pada data multidimensi, tekstual dan gambar, data urutan, dan data spasial. Masalah seperti itu mempelajari manipulasi fitur, biaya kesalahan klasifikasi, dan ketahanan distribusi dalam kesalahan spesifikasi model pembelajaran mendalam. Algoritme pembelajaran mesin yang dihasilkan memiliki aplikasi untuk memodelkan risiko keamanan siber dalam keamanan web, analisis malware, teknik anti-spoofing, penambangan pola langka, klasifikasi tidak seimbang, deteksi contoh di luar distribusi, penyimpangan konsep, dan penambangan motif. Fungsi kerugian manipulasi dan prosedur pelatihan terkait berlaku untuk evaluasi kelayakan penerapan pembelajaran mendalam. Mereka dapat mensimulasikan perlindungan, risiko, dan tantangan keamanan dunia maya seperti optimasi komputasi dan masalah inferensi statistik. Menghasilkan dan menjelaskan manipulasi data yang merugikan memungkinkan studi tentang efek bias algoritmik dalam pembelajaran mendalam. Lebih lanjut, hal ini dapat menjadi saluran bagi teori pengoptimalan yang kuat yang dikembangkan seputar pembelajaran mesin adversarial.

Dalam pemrosesan gambar dan visi komputer, sumber data memiliki aplikasi dalam forensik gambar untuk mendeteksi gambar yang dimanipulasi dalam intelijen strategis. Pertanyaan tentang asal usul gambar yang mencurigakan menjadi menonjol seiring dengan meningkatnya deepfake di Internet. Deepfakes adalah jaringan pembelajaran mendalam yang mampu menghasilkan berita palsu dan bukti palsu di Internet. Mereka menjadi perhatian publik di media sosial online dan lanskap mesin pencari. Ancaman misinformasi akibat data palsu dapat mencoba menargetkan mata rantai terlemah dalam rantai informasi untuk tujuan pemalsuan. Mereka dapat digunakan untuk memanipulasi opini publik selama pemilu, mendiskreditkan, dan memeras masyarakat. Pendekatan baru untuk mengenali media sintetis harus dimasukkan ke dalam ekstensi browser dan perangkat analisis. Serangan adversarial lebih lanjut terhadap pengacak objek dalam gambar dapat mengganggu hasil visi komputer yang meningkatkan tingkat kesalahan klasifikasi dan menyebarkan informasi palsu. Deepfakes memiliki aplikasi dalam seni kreatif, periklanan, produksi film, dan video game. Hal ini dapat mempengaruhi politik pembuktian yang melibatkan manipulasi audiovisual dalam keterangan saksi. Hal ini dapat dikontekstualisasikan, ditafsirkan ulang, dan disiarkan di Internet. Penelitian Nguyen dkk. memberikan survei tentang model pembelajaran mendalam untuk membuat konten deepfake. Sedangkan Carlini dkk. memiliki pendekatan untuk membangun batas ketahanan dalam jaringan saraf yang harus menghadapi contoh adversarial yang dirancang untuk menyesatkan klasifikasi gambar untuk mengambil tindakan yang tidak diinginkan.

Mengintegrasikan asal data ke dalam pembelajaran mesin akan menciptakan metode penemuan pengetahuan yang kuat, terukur, dan dapat digeneralisasikan yang dapat mendukung keaslian media digital untuk memperoleh hasil yang akurat dan andal dalam kecerdasan yang ditambah dengan pelatihan adversarial. Kecerdasan buatan yang dapat

dijelaskan adalah bidang penelitian yang memajukan pembelajaran mesin dalam forensik media digital dan teknologi prediktif. Model yang kuat untuk pengambilan keputusan berdasarkan data dalam pembelajaran mesin mengasumsikan tersedianya informasi yang tidak sempurna untuk mempelajari parameter sistem dan mengoptimalkan distribusi probabilitas pada data yang tidak pasti dan estimasi yang salah. Dalam pengoptimalan yang kuat, variabel acak yang mendasari fitur pembelajaran mesin dimodelkan sebagai parameter tidak pasti yang termasuk dalam kumpulan ketidakpastian cembung dan pengambil keputusan melindungi sistem pembelajaran mesin dari kasus terburuk dalam kumpulan tersebut. Tujuan optimasi berbasis data kemudian menggunakan observasi variabel acak sebagai masukan pelatihan untuk masalah pemrograman matematika.

Pengambilan keputusan yang kuat melibatkan pemrograman stokastik dan optimasi di bawah batasan probabilistik. Masalah optimasi yang kuat juga dapat dipelajari sebagai masalah penghindaran risiko dengan ukuran risiko empiris untuk rekayasa fitur dengan ketahanan yang berlawanan. Pemodelan generatif mendalam dari manipulasi data adversarial menyelidiki ketergantungan antara pemodelan generatif dan atribusi sebab akibat dalam variabel laten. Ini memiliki aplikasi dalam tugas visi komputer yang bertindak sebagai saluran kontrol dalam sistem fisik di mana tantangan utama dalam menghasilkan gangguan fisik yang kuat adalah variabilitas lingkungan.

1.1 KERANGKA PEMBELAJARAN MESIN

Model pembelajaran mesin tradisional mengasumsikan sampel data pelatihan, sampel data pengujian, dan sampel data validasi mengikuti distribusi data yang sama, independen, dan terdistribusi secara identik. Asumsi ini menciptakan kerentanan keamanan dalam model pembelajaran mesin yang dapat diserang oleh musuh cerdas dengan niat jahat. Mengingat sampel data pelatihan, musuh tersebut merancang contoh adversarial untuk meningkatkan kesalahan model. Mengamankan sistem pembelajaran dari contoh-contoh adversarial tersebut merupakan bidang penelitian aktif dalam kecerdasan buatan, diagnostik keamanan, pembelajaran generatif, pembelajaran mendalam, keamanan informasi, sistem otonom, sistem cerdas, dan analisis data.

Contoh manipulasi dapat menyesatkan model pembelajaran selama serangan musuh direncanakan setelah model pembelajaran menyelesaikan pelatihan dan oleh karena itu tidak dapat bereaksi terhadap sampel baru. Dari pengamatan ini, algoritma adversarial memasukkan adversarial ke dalam proses pelatihan model pembelajaran. Dengan demikian, algoritma adversarial memodelkan manipulasi pembelajaran mesin sebagai interaksi antara dua agen model pembelajaran dan satu atau lebih musuh yang cerdas. Teori permainan memberikan kerangka kerja untuk mempelajari interaksi antara model pembelajaran (atau disingkat pembelajar) dan musuh yang cerdas (atau disingkat musuh) dalam hal interaksi antara strategi yang berkembang dari pembelajar dan musuh. Interaksi teori permainan pertama kali dirumuskan dalam ilmu kehidupan sebagai persamaan diferensial non-linier yang mempelajari interaksi antar populasi sistem biologis.

Dalam pembelajaran mesin, fungsi kerugian mengukur dampak ketidakpastian informasi terhadap distribusi prediksi analitik. Algoritma adversarial memformulasikan fungsi kerugian pembelajaran mesin untuk proses pelatihan yang mencegah model overfitting pada data pelatihan di hadapan musuh yang rasional dan adaptif yang menyimulasikan perubahan yang berkembang pada lingkungan pembelajaran sebagai contoh adversarial. Dalam pembelajaran adversarial teoretis permainan, contoh adversarial dihasilkan dengan merancang algoritme pembelajaran mesin dalam berbagai skenario serangan di ruang strategi musuh. Strategi serangan optimal untuk manipulasi adversarial dirumuskan sebagai solusi untuk masalah optimasi (seringkali non-linier dan non-cembung).

Contoh adversarial sulit dideteksi karena model pembelajaran mesin yang dilatih pada data terbatas diperlukan untuk menghasilkan keluaran yang diharapkan untuk setiap masukan yang mungkin. Agen pembelajaran penguatan juga dapat dimanipulasi dengan contoh-contoh adversarial untuk mengakibatkan penurunan kinerja agen di hadapan gangguan yang terlalu halus untuk dirasakan oleh manusia. Dalam tinjauan literatur berikut, kami memberikan gambaran umum tentang algoritme pembelajaran mesin adversarial yang ada, yang masing-masing memiliki skenario serangan dan mekanisme pertahanan yang berbeda untuk menerapkan sistem analisis data yang andal dan sistem pengenalan pola yang kuat. Kami juga merangkum teknik-teknik canggih dalam pembelajaran adversarial teoretis permainan dan pembelajaran penguatan adversarial untuk inferensi dan pengambilan keputusan berbasis perangkat lunak.

Perbandingan Algoritma Manipulasi Pembelajaran Mesin

Bagian ini menyajikan tinjauan literatur dan taksonomi serangan dari algoritma pembelajaran adversarial. Algoritma adversarial dirangkum dalam Tabel 1.1 dalam hal desain algoritma dan penerapan algoritma. Algoritme ini terutama dibandingkan pada fungsi biaya adversarial (atau disingkat fungsi biaya). Ini adalah ukuran kinerja yang diharapkan dari algoritma pembelajaran di hadapan musuh. Ini dirumuskan secara berbeda untuk algoritma pembelajaran adversarial yang berbeda. Kolom tabel mencantumkan berbagai fitur untuk membandingkan algoritma pembelajaran adversarial. Algoritme kami disebut “teori permainan: pembelajaran mendalam”. Baris tabel mencantumkan berbagai algoritma yang dibandingkan. Di seluruh baris, kami mencantumkan model komputasi yang rentan terhadap data adversarial untuk ekstraksi fitur, pembelajaran mendalam, mesin vektor dukungan, dan ansambel pengklasifikasi yang mana data masukan untuk simulasi serangan adversarial dianggap sebagai spam teks, spam gambar, dan spam biometrik.

Algoritma-algoritma tersebut dibandingkan pada fungsi biaya, algoritma pencarian, kondisi konvergensi, strategi serangan, pengaruh serangan, pelanggaran keamanan, pengetahuan musuh, pergerakan algoritma, dan permainan pembelajaran. “*Fungsi biaya*” adalah fungsi tujuan untuk menyelesaikan data adversarial. “*Algoritma pencarian*” adalah algoritma yang digunakan untuk menemukan solusi optimal. “*Kondisi konvergensi*” adalah kriteria pencarian untuk menghasilkan data yang berlawanan. “*Strategi serangan*” adalah skenario serangan dimana musuh beroperasi. “Pengaruh serangan” suatu strategi menentukan akses yang dimiliki musuh untuk melatih data dan menguji input data ke

algoritma pembelajaran. “Pelanggaran kea{manan” adalah tujuan serangan musuh. “*Pengetahuan musuh*” adalah informasi semantik dari musuh. “*Pergerakan algoritma*” adalah tindakan yang diambil oleh algoritma pembelajaran untuk beradaptasi terhadap manipulasi data yang merugikan. Dari tabel tersebut, kita melihat bahwa sebagian besar penelitian yang ada tidak menambahkan formulasi teori permainan ke dalam fungsi biaya. Dengan demikian sebagian besar algoritma pembelajaran yang ada tidak dapat beradaptasi dengan manipulasi data adversarial yang terus menerus. Seperti yang ditunjukkan pada kolom “Game pembelajaran”, ini adalah satu-satunya algoritme pembelajaran adversarial yang memiliki formulasi input distribusi data pelatihan dan pengujian berbasis teori permainan ke model pembelajaran mendalam.

Tabel 1.1 Perbandingan algoritma Adversarial

Algoritma Adversarial	Fungsi	Algoritma Pencarian	Kondisi Konvergen	Strategi Penyerangan	Pengaruh Serangan	Pelanggaran Keamanan	Pengetahuan Musuh	Algoritma bergerak	Permainan Pembelajaran
Pengklasifikasian ensamble	Persamaan 1.1	Pengambilan sampel secara acak	Fitur Ukuran Ansambel, ukuran subset	Susun ulang fitur berdasarkan kepentingan untuk fungsi diskriminan	Kausatif	Ditargetkan, Ketersediaan	Fitur Pelatihan	Fungsi Diskriminasi rata-rata	Tidak ada
Pembobotan Fitur	Perasmaan 1.2	Bagging Fitur	Jumlah Model Dasar	Penambahan / Penghapusan fitur Biner	Kausatif	Menyeluruh, Ketersediaan	Fitur Pelatihan	Perkiraan bobot rata-rata	Tidak ada
Input SVM	Persamaan 1.3	Ascent Gradien	Perubahan Kesalahan Pengujian	Melatih Injeksi Kebisingan	Kausatif	Tertarget, Integritas	Gradien Kerugian	Pembelajaran SVM tambahan	Tidak Ada
Pelabelan SVM	Persamaan 1.4	Ascent Gradien	Suppot Vector Margin (SVM) oleh lp dan qp	Label Injeksi kebisingan	Kausatif	Tertarget, Integritas	Label Pelatihan	Memperbarui bobot SVM hyperplanes	Tidak Ada
Deep Learning	Persamaan 1.5	Propagasi Mundur dengan L-BFGS	Penghentian awal pada kesalahan set validasi adversarial	Gangguan linier pada x	Kausatif	Tertarget, Integritas	Data pelatihan dan Pengujian	Perbarui parameter fungsi keputusan	Tidak ada
Adversarial Network (DNN)	Persamaan 1.6	Augmentasi kumpulan data berbasis Jacobian	Penghentian awal pada kesalahan set validasi adversarial	Amati keluaran DNN Berdasarkan masukan yang dipilih musuh	Penyelidikan	Tertarget, integritas	Menguji Data	Regulasi dan distilasi berbasis Jacobian	Tidak ada
Adversarial Network (DAE)	Persamaan 1.7	Menumpuk DAE kedalam Feed Forward Neural Network	Kesalahan pelatihan	Aditif gaussian loss	Penyelidikan	Tanpa pandang bulu, available	Data Testing	Fungsi penalti menghaluskan data adversarial	Tidak ada
Teori Permainan: Regulalisasi kekurangan	Persamaan 1.8	Wilayah Kepercayaan	Keuntungan adversarial tidak bertambah atau jumlah iterasi maksimum tercapai	Pindahkan data positif ke sampel negatif	Pindahkan data positif ke sampel negatif	Ditargetkan, available	Data Training dan Data Testing	Bangun kembali pengklasifikasian melalui fungsi kerugiannya yang diatur	Permainan Jumlah Nol
Teori Permainan: Serangan yang Jarang	Persamaan 1.9	Anggaran Minimal	Akumulasi biaya melebihi anggaran umum	Fitur tertentu mempengaruhi perkiraan bobot	Kausatif	Tanpa pandang bulu, privasi	Perkiraan bobot	Regulasi parameter	Permainan bukan Jumlah nol
Teori Permainan: SVM	Persamaan 1.10	Pemrograman kuadratik	Kesalahan Pelatihan Tuntuk pada	Hapus fitur berbeda dari titik data berbeda	Kausatif	Tertarget, Integritas	Fitur pelatihan	Perbarui perkiraan bobot untuk	Permainan bukan jumlah nol

			Ketentuan regulasi					manipulasi adversarial	
Teori Permainan: Pembelajaran mendalam (Metode kita)	Persamaan 1.11	Algoritma evousioner	Nash Equilibrium	Pindahan sampel positif ke sampel negatif	Kausatif	Tertarget, Integritas	Data Training	Perbarui perkiraan bobot untuk manipulasi adversarial	Permainan Jumlah Konstan

Persamaan 1.1	$E = 2 - \frac{2}{n}\sum_{k=1}^{n} F(k) = \frac{\sum_{i=1}^{k}\lvert w_i\rvert}{\sum_{j=1}^{n}\lvert w_j\rvert}$
Persamaan 1.2	$min_w \frac{\lambda}{2} w^T w + \frac{1}{m}\sum_{i=1}^{m} l(w^T(S^{-1}x), y),$ $l(f, y) = \max(0, 1 - yf).$
Persamaan 1.3	$max_{x_c} L(X_c) = \sum_{k=1}^{m} (1 - y_k f x_c((x_k))$ $\frac{\partial L}{\partial u} = \sum_{k=1}^{m} M_k \frac{\partial Q_{sc}}{\partial u} + \frac{\partial Q_{kc}}{\partial u}\alpha_c$
Persamaan 1.4	$L(D_{tr}) = argmin_{f\in F}\left[\Omega(f) + C.\hat{R}(f, D_{tr})\right],$ $\hat{R}(f, D_{tr}) = \frac{1}{n}\sum_{i=1}^{n} l(f(x_i), y_i), C > 0,$ $V_L(D_{tr}, D_{vd}) = f_{Dtr}^2 + C.\hat{R}\left(f_{D_{tr}}, D_{vd}\right),$ $f_{D_{tr}} = L(D_{tr}), V_L(z, y) = V_L((x_i, z_i), (x_i, y_i))$
Persamaan 1.5	$\hat{J}(\theta, x, y) = \alpha J(\theta, x, y)$ $+ (1 - \alpha) J\left(\left(\theta, x + \epsilon\, sign\left(\nabla_x J(\theta, x, y)\right)\right)\right).$
Persamaan 1.6	$S_{p+1} = x + \lambda_{p+1}\, sign\left(J_F\left[\hat{O}(x)\right]\right) US_p,$ $\hat{O}(x) = argmax_{j=0\ldots N-1} O_j(x),$ $\delta_x = \epsilon\, sign\left(\nabla_x c(F, x, y)\right),$ $c(F, x, y) = p(y = 1\lvert x) = \exp\left((x - \mu)^T \beta (x - \mu)\right),$ $F(x) = F_n\left(\theta_n, f_{n-1}\left(\theta_{n-1}, \ldots f_2\left(\theta_2, f_1(\theta_1, x)\right)\right)\right).$
Persamaan 1.7	$J_{DCN}(\theta) = \sum_{i=1}^{m}\left(L\left(t^i, y^i\right)\right) + \sum_{j=1}^{H+1}\lambda_j \left\lVert \frac{\partial h_j^i}{\partial h_{j-1}^i}\right\rVert_2$ $min_r c\lvert r\rvert_2 + L(x + r, l).$
Persamaan 1.8	$max_\alpha min_\omega - \lambda\alpha^T\alpha + \lambda\omega^T\omega + C\sum_{i=1}^{n}$ $Loss\left(y_i, w^T(x_i + \alpha)\right).$
Persamaan 1.9	$min_{\omega\epsilon R^d}\lvert y - Xw\rvert_2 + \sum_{i=1}^{d}\lambda_i\lvert w_i\rvert_1$

Persamaan 1.10	$min\frac{1}{2}\|w\|^2 + C\sum_i [1 - y_i w^T x_i + t_i]$ $t_i \geq K_{z_i} + \sum_j v_{ij}$ $v_i \geq 0$ $z_i + v_i \geq (y_i x_i w)$
Persamaan 1.11	$Maxmin: (\alpha^*, w^*) = argmax_{\alpha \epsilon A}$ $J_L = \big(\alpha, argmin_{w \epsilon W} J_L(\alpha, w)\big).$

1.2 MEKANISME KEAMANAN YANG BERLAWANAN

Selain Tabel 1.1, algoritma pembelajaran manipulatif yang ada dan domain penerapannya juga dapat diklasifikasikan berdasarkan mekanisme pertahanan pembelajar dan skenario serangan musuh yang sesuai. Mekanisme pertahanan pelajar telah diusulkan dengan merancang algoritma pembelajaran yang aman, sistem pengklasifikasi ganda, pembelajaran mesin yang menjaga privasi, dan penggunaan pengacakan atau disinformasi untuk menyesatkan musuh.

Biggio dkk. membahas mekanisme pertahanan pelajar dalam kerangka empiris yang memperluas pemilihan model dan langkah-langkah evaluasi kinerja klasifikasi pola oleh Duda et al. Kerangka kerja ini merekomendasikan pelatihan pelajar tentang "keamanan yang dirancang" daripada "keamanan karena ketidakjelasan." Kerangka kerja ini merekomendasikan langkah-langkah tambahan berikut ini untuk memvalidasi mekanisme pertahanan yang diusulkan jika model pembelajaran generatif dan model pembelajaran diskriminatif diserang.

- Secara proaktif mengantisipasi serangan musuh yang paling relevan melalui analisis bagaimana-jika yang menyimulasikan skenario serangan potensial.
- Tentukan skenario serangan berdasarkan tujuan, pengetahuan, dan kemampuan musuh.
- Mengusulkan model distribusi data generatif pada probabilitas bersyarat yang secara formal dapat memperhitungkan sejumlah besar potensi serangan dan sampel validasi silang pada data pelatihan dan data pengujian.

Asumsi berikut dibuat mengenai keamanan algoritma pembelajaran. Kinerja model kemudian dievaluasi berdasarkan strategi serangan optimal yang disimulasikan sesuai dengan kerangka yang diusulkan oleh Biggio et al.

- Sasaran musuh dirumuskan sebagai optimalisasi fungsi sasaran. Fungsi objektif dirancang berdasarkan pelanggaran keamanan yang diinginkan (yaitu integritas, ketersediaan, atau privasi) dan spesifisitas serangan (dari yang ditargetkan hingga yang tidak pandang bulu).
- Pengetahuan musuh didefinisikan sebagai pengetahuan tentang komponen pengklasifikasi, yaitu, data pelatihan, kumpulan fitur, algoritma pembelajaran, fungsi keputusan dan parameternya, ketersediaan dan umpan balik.

- ❖ Kemampuan musuh didefinisikan sebagai kontrol yang dimiliki musuh terhadap data pelatihan dan data pengujian dengan mempertimbangkan batasan spesifik aplikasi seperti pengaruh serangan (baik kausatif atau eksplorasi), efek pada class prior, pecahan sampel, dan fitur yang dimanipulasi oleh musuh.

Tergantung pada tujuan, pengetahuan, dan kemampuan musuh, asumsi-asumsi ini juga diklasifikasikan berdasarkan pengaruh serangan, pelanggaran keamanan, dan kekhususan serangan.

Pengaruh serangan dapat bersifat kausatif atau eksplorasi. Serangan kausatif memengaruhi data pelatihan dan pengujian. Serangan eksplorasi hanya mempengaruhi data pengujian. Pelanggaran keamanan dapat menargetkan integritas atau ketersediaan atau privasi pelajar. Algoritme pembelajaran mesin yang integritasnya dikompromikan tidak dapat mendeteksi perilaku jahat musuh. Integritas algoritme dengan banyak negatif palsu akan terganggu. Algoritme pembelajaran mesin yang ketersediaannya terganggu menunjukkan penurunan kinerja yang parah bagi pengguna yang sah. Ketersediaan algoritma dengan banyak kesalahan positif akan terganggu. Privasi algoritme yang umpan balik detailnya dipublikasikan juga akan terganggu.

Kekhususan serangan dapat ditargetkan atau tidak pandang bulu untuk serangan yang mempengaruhi prediksi atau tindakan algoritma. Dalam serangan yang ditargetkan, serangan diarahkan hanya pada beberapa contoh data pelatihan atau pengujian. Dalam serangan sembarangan, serangan diarahkan pada seluruh kelas instance atau objek. Algoritme adversarial kami memiliki pengaruh serangan kausatif, pelanggaran keamanan integritas, dan kekhususan serangan yang ditargetkan.

Skenario serangan musuh biasanya berkisar antara (i) penambahan gangguan pada fitur/label, (ii) penambahan/penghapusan fitur/label, (iii) sedikit perubahan atau manipulasi atau gangguan pada distribusi data, dan (iv) sedikit perubahan pada keputusan batasan. Masalah optimasi terkait diselesaikan menggunakan algoritma pencarian dengan metode sampling dan gradien. Metode pengambilan sampel berkisar pada pengambilan sampel tambahan, pengambilan sampel bagging, pengambilan sampel bertumpuk, dan pengambilan sampel acak. Metode gradien berkisar antara metode linier, metode kuadrat, metode cembung, dan metode stokastik. Masalah optimasi ini diselesaikan dengan menemukan solusi optimal lokal yang ditentukan oleh kondisi konvergensi yang berkisar pada (i) jumlah fitur, (ii) jumlah istilah regularisasi, dan (iii) perubahan estimasi kesalahan pada data pelatihan/pengujian.

Algoritme adversarial kami menyebabkan sedikit perubahan pada distribusi data yang disimulasikan dengan optimasi stokastik dan metode pengambilan sampel acak. Masalah pengoptimalan kami menyatu ke dalam solusi yang dihitung pada keseimbangan Nash dalam permainan Stackelberg. Dari sudut pandang musuh, solusi keseimbangannya adalah solusi optimum lokal jika terjadi skenario serangan terburuk dan solusi optimal global jika terjadi skenario serangan terbaik. Kekuatan dan relevansi skenario serangan kami ditentukan oleh performa model pembelajaran mendalam yang diserang.

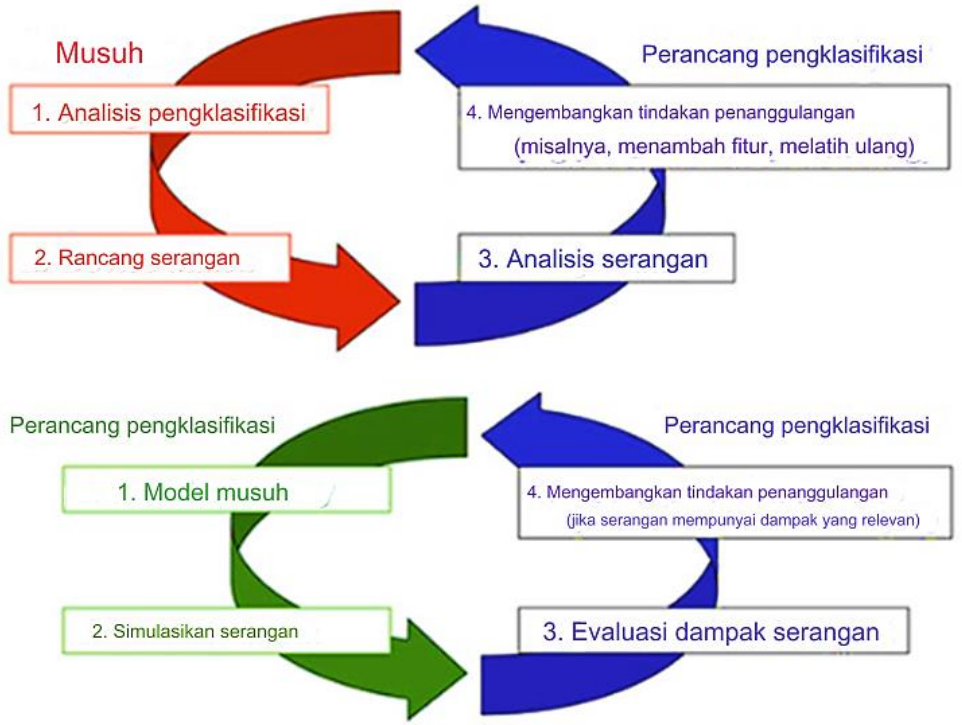
Contoh Manipulasi pembelajaran Taksonomi

Papernot dkk. memberikan model ancaman yang merangkum berbagai skenario serangan dalam algoritma pembelajaran adversarial. Mekanisme pertahanan pengklasifikasi adversarial diharapkan dapat meningkatkan ketahanan model terhadap sampel data validasinya. Di sini, sampel data validasi disebarkan ke dalam distribusi data runtime model terlatih menjadi non-iid sehubungan dengan sampel data pengujian dalam distribusi data pelatihan model terlatih.

Papernot dkk. mengungkapkan model ancaman pembelajaran mesin mereka dalam langkah-langkah manipulasi adversarial yang ditemukan selama proses pelatihan pembelajaran mesin dan proses inferensi pembelajaran mesin. Selama proses pelatihan pembelajaran mesin, musuh seharusnya memanipulasi proses pengumpulan data online atau proses pengumpulan data offline. Manipulasi adversarial semacam itu memasukkan contoh-contoh adversarial atau memodifikasi data pelatihan dengan tujuan mengubah batasan keputusan model pembelajaran. Selama proses inferensi pembelajaran mesin, musuh seharusnya merencanakan serangan blackbox atau serangan whitebox pada parameter model pembelajaran. Pengaturan serangan seperti itu menyebabkan penyimpangan distribusi antara distribusi data pelatihan dan waktu proses.

Papernot dkk. juga melihat keamanan pembelajaran mesin melalui prisma model kerahasiaan, integritas, dan ketersediaan di mana musuh masing-masing menargetkan parameter, label, dan fitur pengklasifikasi. Berbeda dengan keamanan pembelajaran mesin, privasi pembelajaran mesin dieksplorasi dalam hal performa model (i) distribusi data pelatihan dan waktu proses berbeda, (ii) jumlah data yang diekspos oleh model pembelajaran terikat oleh anggaran privasi yang berbeda, dan (iii) pertahanan model pembelajaran memberikan keadilan, interpretasi, dan transparansi terhadap keluaran pembelajaran. Lingkungan adversarial yang memengaruhi kompleksitas model, akurasi model, dan ketahanan model dirumuskan dalam teorema tidak ada makan siang gratis untuk pembelajaran adversarial. Papernot dkk. juga memotivasi pembelajaran adversarial teoretis permainan selama inferensi pembelajaran mesin dalam kerangka pembelajaran yang mungkin kurang lebih benar (PAC).

Biggio dkk. mensurvei pembelajaran mesin adversarial untuk pengklasifikasi pola. Contoh adversarial untuk pengklasifikasi pola seharusnya dibuat pada waktu pelatihan atau waktu pengujian. Penelitian terbaru mengenai contoh adversarial untuk aplikasi jaringan dalam dalam visi komputer dan keamanan siber juga dibahas. Skenario serangan pada waktu pelatihan disebut serangan keracunan, sedangkan skenario serangan pada waktu pengujian disebut serangan penghindaran. Untuk diintegrasikan dengan terminologi pembelajaran mendalam, serangan keracunan juga disebut serangan pelatihan manipulasi pembelajaran , sedangkan serangan penghindaran juga disebut serangan pengujian manipulasi pembelajaran . Kemudian evaluasi keamanan dan mekanisme pertahanan pengklasifikasi pola yang diserang dibahas. Di sini juga disajikan model pembelajaran proaktif security-by-design yang menggabungkan desain musuh dalam proses pembelajaran. Hal ini ditunjukkan pada Gambar 1.1.



Gambar 1.1 Perlombaan senjata reaktif dan proaktif antara musuh dan pembelajar

Biggio dkk. mengategorikan desain musuh sebagai pemecahan masalah pengoptimalan untuk strategi serangan terbaik yang ditentukan oleh tujuan musuh dalam skenario serangan, pengetahuan musuh tentang sistem pembelajaran yang ditargetkan, dan kemampuan musuh dalam memanipulasi data masukan. Berdasarkan berbagai asumsi pada desain musuh tersebut, strategi serangan yang optimal kemudian terbukti memungkinkan tidak hanya untuk algoritma pembelajaran yang diawasi tetapi juga algoritma pembelajaran tanpa pengawasan seperti algoritma pengelompokan dan algoritma pemilihan fitur. Sasaran Musuh selanjutnya dikategorikan menjadi (i) pelanggaran keamanan yang membahayakan integritas, ketersediaan, dan privasi sistem pembelajaran dan (ii) spesifisitas serangan dan spesifisitas kesalahan yang menyebabkan kesalahan klasifikasi pada kumpulan sampel tertentu dan kumpulan kelas tertentu. Di sini, pengetahuan musuh tentang sistem pembelajaran yang ditargetkan dikategorikan lebih lanjut menjadi berikut

- Serangan whitebox pengetahuan sempurna dengan pengetahuan lengkap tentang parameter pembelajaran. Dalam hal ini, evaluasi keamanan memberikan batasan atas penurunan kinerja dalam skenario serangan.
- Serangan kotak abu-abu dengan pengetahuan terbatas dengan pengetahuan sebelumnya tentang representasi fitur dan algoritma pembelajaran tetapi tidak data pelatihan dan parameter pembelajaran. Di sini evaluasi keamanan dilakukan pada pengklasifikasi pengganti yang mempelajari kumpulan data pengganti yang tersedia dari sumber data serupa dengan data pelatihan. Contoh adversarial untuk pengklasifikasi pengganti kemudian diuji terhadap pengklasifikasi yang ditargetkan

untuk mengevaluasi kemampuan transfer skenario serangan antar algoritma pembelajaran.

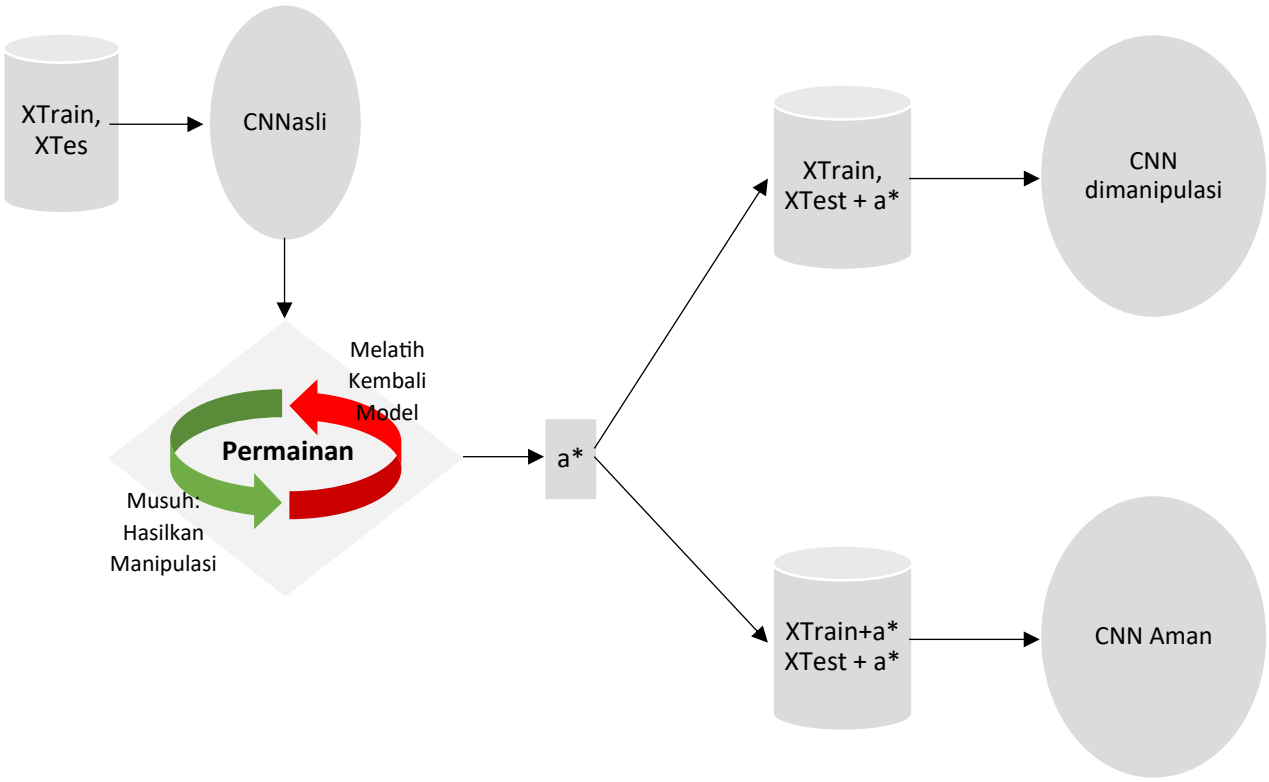
- Serangan blackbox tanpa pengetahuan tanpa pengetahuan apa pun tentang algoritma pembelajaran, namun pengetahuan parsial tentang representasi fitur dan distribusi data pelatihan. Di sini evaluasi keamanan memeriksa apakah strategi serangan optimal ditransfer antara model pengganti yang terlatih secara optimal dan model pengklasifikasi yang ditargetkan. Umpan balik yang diperkuat pada keputusan pengklasifikasi dapat digunakan untuk menyempurnakan model pengganti.

Biggio dkk. juga mengkategorikan pengetahuan musuh berdasarkan batasan manipulasi data spesifik aplikasi pada distribusi data input, fitur, dan kelas. Formulasi tingkat tinggi mengenai strategi serangan optimal musuh dan kurva evaluasi keamanan pengklasifikasi juga disediakan.

Evaluasi keamanan semacam itu mempertimbangkan algoritma pembelajaran yang dapat dibedakan dan tidak dapat dibedakan seperti jaringan saraf dan pohon keputusan. Di sini analisis sensitivitas jaringan dalam didefinisikan sebagai studi tentang fenomena gangguan minimal sampel pelatihan, sedangkan evaluasi keamanan yang lebih umum dari pengklasifikasi pola didefinisikan sebagai studi tentang kekuatan serangan musuh dan kepercayaan diri serangan dalam memanipulasi batasan keputusan pengklasifikasi untuk target. kelas. Teknik pertahanan proaktif yang dirangkum dalam rangkaian serangan tersebut mencakup (i) pengacakan data pelatihan dan keluaran pengklasifikasi, (ii) pakar domain yang mengoreksi keputusan pengklasifikasi, (iii) sanitasi data dengan statistik yang kuat, (iv) deteksi penyimpangan otomatis, (v) dengan benar menggabungkan ansambel pengklasifikasi, (vi) heuristik pelatihan adversarial berulang, (vii) pembelajaran adversarial teoretis permainan, dan (viii) pengoptimalan yang kuat dalam pembelajaran teratur yang secara efektif mengatasi kutukan dimensi pada kumpulan data besar dan pengklasifikasi non-linier. Penelitian di masa depan tentang evaluasi keamanan pembelajaran mesin adversarial berbasis data diusulkan untuk dilakukan pada titik temu antara pengujian perangkat lunak, verifikasi formal, kecerdasan buatan yang kuat, dan pembelajaran mesin yang dapat ditafsirkan.

1.3 ILUSTRASI PERMAINAN STOKASTIK DALAM MENGELOLA PEMBELAJARAN MESIN

Gambar 1.2 menggambarkan proses pembelajaran pada formulasi permainan penelitian kami dalam bentuk diagram alir. CNNoriginal dilatih pada data pelatihan Xtrain dan dievaluasi pada data pengujian Xtest untuk memberikan “kinerja pelajar” dalam eksperimen. Gambar 1.2 mengilustrasikan permainan dua pemain. Permainan ini memiliki gerakan yang dilakukan oleh masing-masing lawan dan pembelajar selama setiap interaksi. Dalam gerakan ini, musuh menargetkan pelajar dengan sampel adversarial yang dihasilkan dari operator evolusi. Pelajar kemudian mengadaptasi operator pembelajaran mendalam untuk data adversarial dengan melatih ulang CNN pada sampel validasi silang yang baru.



Gambar 1.2 Diagram alur yang mengilustrasikan manfaat pembelajar teori permainan. Permainan dua pemain dimainkan oleh satu musuh dan satu pembelajar. Game ini menghasilkan jaringan pembelajaran mendalam akhir CNNsecure yang lebih siap menghadapi manipulasi adversarial dibandingkan jaringan pembelajaran mendalam awal CNNoriginal

Sekumpulan L dari M musuh L L1, L2, L3,..., LM menargetkan kinerja ini dengan melibatkan CNN dalam beberapa permainan berurutan dua pemain. Dalam setiap permainan dua pemain, CNN yang dilatih pada sampel data asli dan yang dihasilkan serta diuji pada data adversarial masing-masing adalah CNN yang dimanipulasi cnn dan CNN yang dimanipulasi gan. Semua CNN ini diberikan di bawah istilah umum “pebelajar yang dimanipulasi pertunjukan." Kami menemukan bahwa CNNmanipulated−cnn serta CNNmanipulated−gan memiliki performa yang jauh lebih buruk dibandingkan CNN asli CNN asli yang dilatih pada data pelatihan dan pengujian asli (Xtrain, Xtest). Dengan demikian kami menyimpulkan manipulasi adversarial berhasil menyerang pembelajar. Jaringan neural konvolusional baru CNNsecure kemudian dilatih ulang (Xtrain AS∗ , Xtest AS∗) untuk beradaptasi dengan manipulasi adversarial. Ini diberikan sebagai “kinerja pelajar yang aman.” CNNsecure adalah model yang kami usulkan. Hal ini ditemukan lebih baik daripada CNNmanipulated−cnn dan CNNmanipulated−gan yang dimanipulasi.

Oleh karena itu, kami menyimpulkan bahwa CNNsecure baru telah berhasil beradaptasi dengan data adversarial yang dihasilkan oleh banyak musuh, sedangkan CNN asli yang diberikan rentan terhadap setiap manipulasi adversarial αi∗ yang dihasilkan oleh setiap musuh Li yang memainkan permainan i pada distribusi data pelatihan/pengujian yang diberikan. Algoritma kami mampu menemukan sampel data yang mempengaruhi kinerja

CNN. CNN yang mampu pulih dari serangan musuh kami lebih siap menghadapi perubahan tak terduga dalam distribusi data mendasar. Permainan antara musuh dan pembelajar memungkinkan kita menghasilkan manipulasi data adversarial untuk CNN yang dilatih tentang distribusi data yang mendasarinya.

BAB 2
MANIPULASI PEMBELAJARAN MENDALAM

Pembelajaran mendalam terbukti tidak aman. Jaringan saraf dalam rentan terhadap serangan keamanan dari musuh jahat, yang merupakan tantangan berkelanjutan dan penting bagi para peneliti pembelajaran mendalam. Bab ini mempelajari algoritma pembelajaran mendalam yang bermusuhan dalam mengeksploitasi kerentanan jaringan saraf dalam. Fokus utamanya adalah pada serangkaian algoritma pembelajaran mendalam adversarial game teoritis untuk meningkatkan ketahanan jaringan terutama dalam skenario serangan kotak hitam tanpa pengetahuan. Meskipun ada banyak karya terbaru yang mempelajari kerentanan jaringan, hanya sedikit yang mengusulkan serangan kotak hitam tanpa pengetahuan, dan bahkan lebih sedikit lagi yang menggunakan pendekatan berbasis teori game. Bahkan gangguan yang tidak berbahaya pada data pelatihan dapat mengubah perilaku jaringan dalam dengan cara yang tidak diinginkan.

Artinya, penyimpangan kecil yang tidak terlihat dan tidak terukur pada data pelatihan dapat menghasilkan klasifikasi label yang sangat berbeda saat menggunakan model untuk pembelajaran mendalam yang diawasi. Detail algoritmik yang diusulkan dalam bab ini telah digunakan dalam pembelajaran mendalam teori permainan dengan musuh evolusioner, musuh stokastik, musuh acak, dan musuh variasional yang diusulkan dalam penelitian kami. Dalam merancang skenario serangan, tujuan adversarial adalah membuat perubahan kecil dan tidak terdeteksi pada data pengujian. Musuh memanipulasi parameter representasi dalam data masukan untuk menyesatkan proses pembelajaran jaringan saraf dalam, sehingga berhasil salah mengklasifikasikan label kelas asli sebagai label kelas yang ditargetkan.

Deng mensurvei literatur yang ada tentang pembelajaran mendalam untuk pembelajaran representasi dan pembelajaran fitur di mana hierarki fitur atau konsep tingkat yang lebih tinggi ditentukan dari fitur atau konsep tingkat yang lebih rendah. Model, arsitektur, dan algoritma pembelajaran mendalam dikategorikan ke dalam tiga kelas—model generatif, diskriminatif, dan hibrid:

- Model generatif mengkarakterisasi distribusi probabilitas gabungan dari data yang diamati dan kelas terkaitnya dengan sifat korelasi tingkat tinggi antara variabel yang diamati dan variabel tersembunyi.
- Model diskriminatif membedakan pola-pola dengan mengkarakterisasi distribusi kelas-kelas posterior yang dikondisikan pada data yang diamati.
- Model hibrid adalah model diskriminatif yang dibantu oleh model generatif secara signifikan melalui optimalisasi yang lebih baik atau/dan regularisasi kriteria diskriminatif yang digunakan untuk mempelajari parameter dari data.

Pembelajaran mendalam juga dipahami oleh Deng sebagai perpanjangan dari penelitian sebelumnya pada arsitektur dangkal yang memecahkan masalah yang dibatasi dengan baik seperti model linier umum, perceptron multi-layer, mesin vektor pendukung, model entropi

maksimum, bidang acak bersyarat, campuran Gaussian. model, dan model Markov tersembunyi. Untuk masalah umum yang ditangani oleh arsitektur dalam, arsitektur dangkal dan metode statistiknya cenderung menghasilkan algoritma komputasi yang sulit untuk inferensi kelas.

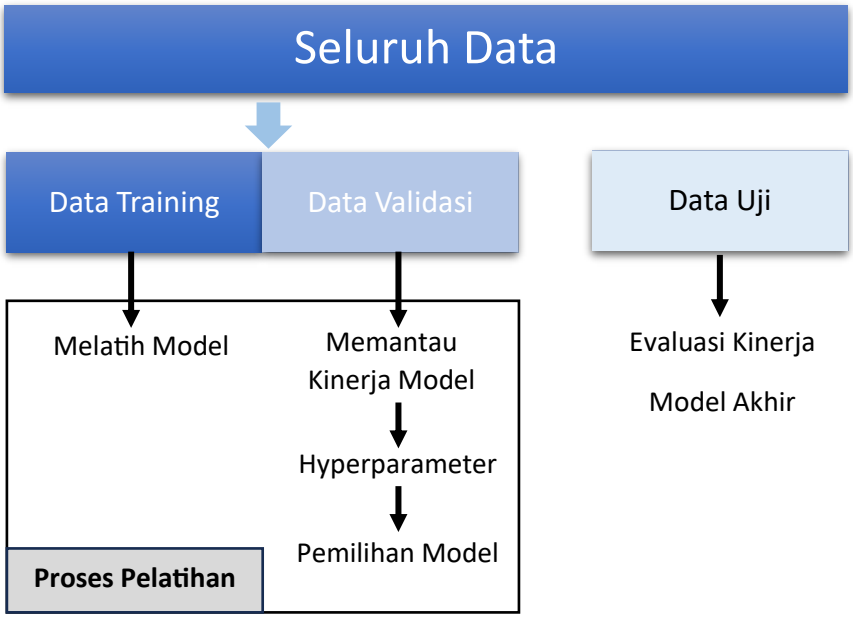
Model pembelajaran mendalam yang umum digunakan seperti jaringan keyakinan mendalam, autoencoder variasional, dan jaringan saraf konvolusional mengekstrak struktur dan keteraturan dalam fitur masukan dengan menghindari kesulitan dalam pengoptimalan global. Pengoptimalan parameter dilakukan dengan merancang algoritme pelatihan lapis demi lapis serakah yang membantu meringankan masalah overfitting yang diamati di banyak arsitektur dangkal yang melatih jutaan parameter. Oleh karena itu, model pembelajaran mendalam berguna untuk pembelajaran end-to-end sistem cerdas yang menanamkan pengetahuan domain dan menafsirkan ketidakpastian.

2.1 ANALISIS KURVA PEMBELAJARAN UNTUK PEMBELAJARAN MESIN YANG DIAWASI

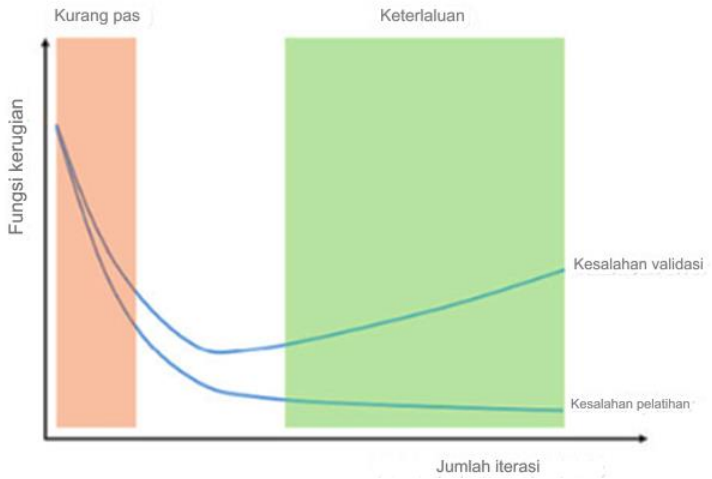
Teorema tidak ada makan siang gratis (NFL) untuk pembelajaran dan pengoptimalan yang diawasi menyatakan bahwa, rata-rata dari semua situasi teoretis pembelajaran yang diwakili dalam sampel data, model pembelajaran mesin yang lebih memilih pelatihan sederhana hingga pelatihan kompleks sering kali gagal karena berhasil. Artinya, proses acak yang menghasilkan distribusi data pelatihan mungkin tidak selalu sama dengan proses acak yang mengatur distribusi data pengujian. Ada banyak model alternatif yang perlu dipertimbangkan untuk analisis data yang bercampur dengan noise. Tidak ada jaminan bahwa model statistik yang dipilih adalah model yang benar atau cukup menangkap pola di seluruh sampel data. Teknik pemulusan dan regularisasi adalah pendekatan sederhana untuk mengungkap pola dalam data pelatihan dengan prasangka dan asumsi minimal mengenai pola apa yang seharusnya ada dalam data pengujian. Secara umum, kita harus bersaing dengan kriteria pemilihan model untuk algoritma analitik yang dipilih.

Dalam model analitik prediktif yang dibangun dengan algoritme pembelajaran mesin yang diawasi, kriteria pemilihan model melakukan optimalisasi kecocokan terhadap sampel data pelatihan dan pengujian. Hal ini disebut validasi silang yang mengasumsikan bahwa model statistik sama bagusnya dengan prediksinya. Skema evaluasi model ini tidak mampu memperkirakan prediksi kontrafaktual ketika dunia berubah. Jadi Gambar 2.1 menunjukkan sampel data validasi tambahan untuk membandingkan kelas prediksi dengan kelas sebenarnya. Dalam pembelajaran adversarial, perbandingan tersebut dilakukan dengan fungsi biaya manipulasi pembelajaran yang memperhitungkan informasi distribusi kelas dan biaya dalam menghasilkan prediksi algoritma pembelajaran yang diawasi. Dengan demikian, data manipulasi pembelajaran dapat dianggap sebagai bagian dari sampel data validasi dalam pemilihan model. Proses pelatihan adversarial melatih model pembelajaran mesin pada sampel data pelatihan yang diberikan oleh pengguna dan sampel data validasi yang dibuat oleh musuh. Selanjutnya, sampel data validasi digunakan untuk menyempurnakan hyperparameter untuk melatih model pembelajaran mesin.

Dalam evaluasi eksperimental pembelajaran mesin adversarial, kita dapat menjalankan uji statistik untuk menemukan skenario kontrafaktual dalam data pelatihan. Inferensi kausal juga dapat digunakan untuk memperkirakan dampak skenario kontrafaktual. Untuk pemilihan model sistematis dalam pembelajaran mesin, fokus pemodelan kontrafaktual adalah memperkirakan apa yang akan terjadi jika terjadi perubahan yang mungkin benar-benar terjadi atau tidak terjadi pada data pelatihan. Model pembelajaran mesin yang bermusuhan seperti itu mungkin mengorbankan kinerja prediktif di lingkungan saat ini agar pembelajaran mesin dapat menemukan fitur kontrafaktual baru dalam lingkungan validasi yang berubah untuk pembelajaran mesin. Kebijakan kontrafaktual yang dihasilkan membandingkan data pelatihan dengan data validasi dapat digunakan untuk menentukan analisis sensitivitas baru, deteksi anomali, dan penerapan penyimpangan konsep untuk pembelajaran adversarial. Metrik evaluasi yang sensitif terhadap biaya memperhitungkan perbedaan tingkat keparahan antara alarm palsu dan kasus penipuan yang terlewat.



Gambar 2.1 Proses pelatihan fungsi kerugian manipulasi pembelajaran



Gambar 2.2 Kurva pembelajaran fungsi kerugian kustom

Teknik pemeringkatan fitur kemudian dapat memandu sinyal kontekstual dari prediksi palsu dan manipulasi fitur. Mereka memperhitungkan tingkat sensitivitas algoritma klasifikasi yang berbeda terhadap fitur palsu dalam sampel data pelatihan. Dengan adanya validasi adversarial, pembelajaran mendalam menunjukkan tingkat konvergensi yang lambat dan sensitivitas terhadap kebisingan. Jadi kita harus membuat kurva pembelajaran pada pembelajaran mendalam seperti pada Gambar 2.2 untuk menemukan fitur kontrafaktual dalam garis dasar klasifikasi berkode warna yang menunjukkan kinerja pada sumbu y untuk rentang parameter pada sumbu x. Berdasarkan trade-off bias-varians dalam pembelajaran mesin, model kompleks yang cenderung menggunakan data noise secara berlebihan menunjukkan varians yang tinggi, sedangkan model sederhana yang kurang fleksibel untuk memperkirakan proses kompleks menunjukkan bias yang tinggi.

Kami ingin mendapatkan kriteria kesesuaian untuk pemilihan model yang tidak underfitting dengan bias tinggi atau overfitting dengan varian tinggi terhadap sampel data pelatihan. Secara praktis, kami ingin memilih wilayah pada Gambar 2.2 yang menunjukkan kesalahan rendah pada semua sampel data pelatihan, validasi, dan pengujian. Overfitting terjadi ketika kesalahan pelatihan rendah tetapi kesalahan pengujian tinggi. Underfitting terjadi ketika kesalahan pengujian rendah tetapi kesalahan pelatihan tinggi. Dalam menganalisis kesalahan prediksi, dekomposisi bias-varians memisahkan analisis bias dan varians dalam evaluasi model pembelajaran mesin. Dengan mem-bootstrap sampel dari data yang diberikan dalam eksperimen validasi silang, kami membuat sampel data pelatihan, validasi, dan pengujian untuk memperkirakan model dari algoritme pembelajaran mesin. Bagging, boosting, dan stacking adalah metode pengambilan sampel data yang umum digunakan untuk membuat kumpulan data validasi silang. Dekomposisi bias varians berlaku untuk kesalahan generalisasi yang dihasilkan dari fungsi kerugian dalam klasifikasi dan regresi.

Kurva pembelajaran mewakili kinerja generalisasi model yang dihasilkan oleh algoritma pembelajaran. Berdasarkan perkiraan probabilitas keanggotaan kelas, kurva pembelajaran membandingkan algoritma klasifikasi yang berbeda untuk mengeksplorasi hubungan antara ukuran dataset pelatihan dan algoritma pembelajaran/induksi. Mereka memungkinkan kita melihat pola yang umum di berbagai kumpulan data. Tanpa pemeriksaan kurva pembelajaran, kita tidak dapat menarik kesimpulan bahwa satu algoritma lebih baik dari algoritma lain untuk domain aplikasi tertentu. Ringkasan analisis kurva pembelajaran diberikan oleh Perlich et al.

Perbandingan antara pemodelan analitik untuk mendapatkan bukti teoretis yang kuat harus dilakukan dengan metrik kinerja pada ketidakseimbangan biaya akibat kesalahan klasifikasi yang salah dalam prediksi. Performa prediktif dari kemampuan model untuk membedakan antara data adversarial dan data pelatihan dapat dianalisis dengan akurat dan berada di bawah kurva operasi penerima (AUC). Selain itu, metrik kinerja yang mencerminkan ketidakseimbangan dalam label kelas dapat digunakan untuk menghitung kesalahan klasifikasi. Itu termasuk sensitivitas, skor F1, dan skor F2. Makalah kurva belajar. Penting untuk memasukkan perlindungan perlindungan data tersebut ke dalam rantai nilai analitik

yang dibangun dengan validasi silang atau mengadakan pengujian untuk memilih algoritma yang “paling akurat” untuk menganalisis kumpulan data tertentu.

Analisis risiko dunia maya untuk kebocoran informasi dalam pembelajaran mendalam menjadi penting untuk menganalisis model pembelajaran mesin yang dilatih pada kumpulan data sensitif. Kurva pembelajaran dapat mempertimbangkan dekomposisi bias-varians dalam fungsi kerugian manipulasi pembelajaran untuk mendapatkan regularisasi dalam tujuan pembelajaran pembelajaran mesin yang diawasi. Selama eksperimen validasi model, perbedaan informasi antara sampel data validasi dan sampel data pelatihan dapat dihitung sebagai distribusi data terdiskritisasi yang diperoleh dari skema pengambilan sampel dalam pembelajaran mendalam adversarial.

Sejauh mana kebisingan pada parameter pemodelan dan data pelatihannya dapat bermanfaat bagi kualitas distribusi data secara keseluruhan dalam skema pengambilan sampel bergantung pada proses kebisingan manipulasi pembelajaran tertentu dan sifat distribusi target yang dihasilkan dalam pembelajaran adversarial teoritis permainan. Proposal kami mempelajari interaksi fungsi biaya manipulasi pembelajaran dan fungsi kesalahan klasifikasi untuk merancang pengklasifikasi teoritis permainan yang memburuk pada tingkat yang lebih lambat dibandingkan pengklasifikasi biasa pada data manipulasi pembelajaran . Dengan pemodelan generatif yang mendalam mengenai strategi respons terbaik dari musuh, kami membangun dinamika pengambilan sampel data dari studi pengukuran terhadap musuh yang sensitif terhadap biaya untuk pembelajaran diskriminatif. Kami mensimulasikan pengkodean batas keputusan yang dihasilkan sebagai masalah penyimpanan-pengambilan dalam penambangan data.

2.2 FUNGSI KERUGIAN PEMBELAJARAN MESIN MANIPULATIF

Contoh manipulasi dapat dibuat dengan bereksperimen pada fungsi kerugian dalam pembelajaran mendalam. Eksperimen semacam itu menghasilkan analisis data empiris seputar fungsi kerugian yang merugikan dan prosedur pelatihan yang sesuai dalam pembelajaran diskriminatif. Penelitian ini kemudian dapat diterapkan pada studi tentang kepercayaan pembelajaran mendalam dalam sistem cyber-fisik dalam penerapan di dunia nyata.

Kita dapat merumuskan dan menyesuaikan fungsi kerugian manipulasi pembelajaran dengan fungsi tujuan pembelajaran adversarial yang diselesaikan dengan algoritma optimasi. Algoritme pembelajaran adversarial kemudian dapat merancang manipulasi data pelatihan dengan target musuh untuk menyesatkan jaringan saraf dalam. Setelah menunjukkan kerentanan pembelajaran mendalam dengan cara ini, kami dapat mengusulkan mekanisme pertahanan untuk membuat jaringan saraf yang kuat. Khusus untuk aplikasi pembelajaran yang diawasi, fungsi kerugian mengevaluasi kesalahan statistik analisis prediktif. Biasanya, fungsi kerugian mengurangi bias dalam model klasifikasi prediktif dan varians dalam model regresi prediktif. Di sini, fungsi kerugian manipulasi pembelajaran mengurangi sensitivitas model prediktif terhadap noise model.

Proposal kami adalah menganalisis jenis kebisingan ini dalam paradigma pembelajaran mendalam adversarial teoretis permainan. Ini melibatkan desain fungsi pembayaran adversarial yang menghasilkan manipulasi data adversarial dengan mengoptimalkan fungsi biaya adversarial untuk berbagai jenis musuh. Musuh tersebut termasuk musuh evolusioner, stokastik, pengacakan, variasional, dan generatif.

Intuisi fungsi kerugian manipulasi pembelajaran kita berasal dari konsep tindakan dan gerakan dalam teori permainan. Selama pembelajaran, skenario serangan dimodelkan sebagai gerakan yang dilakukan oleh algoritma pembelajaran dan gerakan balasan yang dilakukan oleh musuh yang cerdas. Teori permainan kami mempelajari interaksi antara agen atau pemain independen yang bekerja untuk mencapai tujuan. Setiap pemain memiliki serangkaian strategi/gerakan/tindakan terkait yang mengoptimalkan fungsi pembayaran atau fungsi utilitas untuk mencapai tujuan. Permainan pada akhirnya menyatu ke keadaan ekuilibrium dimana tidak ada pemain yang mempunyai insentif untuk menyimpang.

Melalui optimalisasi pembelajaran adversarial yang diusulkan, kita dapat menganalisis secara empiris fungsi kerugian diskriminatif dalam pembelajaran mendalam untuk menghasilkan titik data yang salah klasifikasi dan karenanya melakukan manipulasi adversarial pada data pelatihan. Lebih lanjut, berbeda dengan metode pembelajaran mendalam tradisional, kami mengusulkan fungsi imbalan adversarial yang tidak dapat dibedakan dan terputus-putus dalam ruang pencarian manipulasi adversarial. Dalam kerangka minimalisasi risiko empiris untuk pembelajaran terawasi dan teori permainan, kami mempelajari fungsi kerugian adversarial untuk pembelajaran diskriminatif yang melibatkan klasifikasi dan regresi.

Masalah ilmu data yang kaya dan fitur pembelajaran mesin dapat direkayasa dari algoritme kami dengan memodelkan beragam skenario aplikasi analisis data yang melibatkan pembelajaran diskriminatif. Misalnya, kami mengusulkan fungsi kerugian manipulasi pembelajaran untuk mempelajari momen dan kumulan distribusi data yang bergantung pada waktu dalam pemodelan regresi. Fungsi kerugian yang diusulkan dapat diperluas untuk pendekatan berorientasi algoritma non-linier menuju regresi yang kuat. Sensitivitas fungsi kerugian kami dapat disesuaikan dengan pola yang dibangun untuk meningkatkan pemilihan model yang bergantung pada aplikasi. Di sini, model generatif mendalam berguna untuk rekayasa fitur dan generalisasi pembelajaran dalam domain aplikasi tertentu.

Fungsi imbalan adversarial kami dapat memodelkan hipotesis diskriminasi seputar label kelas dan batasan keputusannya dalam pemodelan klasifikasi. Fungsi pembayaran yang diusulkan mengoptimalkan pencarian manipulasi data pada ruang data piksel asli serta ruang data laten yang mewakili distribusi piksel dengan model campuran Gaussian. Fungsi pembayaran kemudian dioptimalkan dengan pengaturan parameter dalam algoritma simulasi anil, pembelajaran variasional, dan pembelajaran generatif. Kinerja kesalahan klasifikasi jaringan saraf dalam pada masa keseimbangan Nash diukur dalam bentuk statistik-t yang dihipotesiskan atas perolehan, tingkat positif sebenarnya, dan skor f1 dari label kelas yang ditargetkan.

Kami bereksperimen dengan fungsi pembayaran adversarial pada ruang strategi acak dengan mengubah formulasi permainan Stackelberg. Di sini, skenario serangan terhadap ruang strategi menentukan kriteria konvergensi permainan Stackelberg pada kumpulan data multi-label. Dalam ekuilibrium Nash, permainan ini terpusat pada manipulasi adversarial yang memengaruhi kinerja pengujian di seluruh label yang ditargetkan baik dalam model klasifikasi dua label maupun multi-label. Hasilnya mengarahkan kami pada proposal untuk pelajar yang aman dan kebal terhadap jenis serangan adversarial tersebut, dan analisis empiris menegaskan bahwa model klasifikasi ini secara signifikan lebih kuat daripada jaringan saraf dalam tradisional yang diserang oleh musuh.

2.3 CONTOH MANIPULASI DI JARINGAN DALAM

Papernot dkk. menyajikan demonstrasi praktis sampel adversarial yang diketahui berpindah antar model pembelajaran mendalam. Contoh adversarial tersebut dibuat untuk mengontrol integritas jaringan saraf dalam (DNN) target tanpa akses ke arsitektur, parameter, dan data pelatihan DNN target. DNN pengganti kemudian dilatih untuk memperkirakan model target DNN yang dipelajari. DNN pengganti juga tidak memiliki pengetahuan tentang vektor probabilitas yang mengkode keyakinan DNN target tentang hubungan antara masukan pelatihan dan kelas. Serangan ini didefinisikan dengan asumsi bahwa musuh dapat mengamati keluaran DNN target berdasarkan masukan yang dipilih oleh musuh. Model musuh memiliki akses ke distribusi data pelatihan yang sama dengan model target.

DNN pengganti dilatih melalui teknik augmentasi kumpulan data berbasis Jacobian. Langkah dalam algoritma ini disebut pelatihan model pengganti. Teknik augmentasi kumpulan data ini memungkinkan musuh memilih titik data yang mewakili perilaku DNN target di domain masukan. Serangan adversarial dibuat mudah dilakukan dengan membatasi jumlah kueri yang dimasukkan ke DNN target. Kueri dirumuskan oleh heuristik pencarian yang efisien pada domain data masukan.

Setelah menemukan contoh-contoh adversarial, algoritme menyempurnakan gangguan dalam contoh-contoh adversarial untuk memaksimalkan kemampuan transfer sampel adversarial. Penyempurnaan ini didasarkan pada pengamatan bahwa matriks tanda gradien biaya DNN pengganti dan DNN target berkorelasi. Langkah dalam algoritme ini disebut pembuatan sampel manipulasi pembelajaran. Secara keseluruhan, algoritma manipulasi pembelajaran melewati fase pengumpulan awal, pemilihan arsitektur, pelabelan, pelatihan, dan augmentasi. Algoritma manipulasi pembelajaran menghasilkan set pelatihan pengganti yang mewakili batasan keputusan model target.

Dua algoritma diimplementasikan untuk mencari sampel adversarial. Kedua algoritme penelusuran mengevaluasi sensitivitas model target terhadap masukan model pengganti sehingga gangguan kecil pada masukan model target mencapai tujuan kesalahan klasifikasi yang berlawanan. Kedua algoritma pencarian berbeda dalam efisiensi komputasi dalam menghasilkan contoh yang berlawanan. Pertahanan yang diusulkan terhadap serangan tersebut mencakup penggunaan pelatihan sampel adversarial, regularisasi dan distilasi berbasis Jacobian, dan analisis yang cermat terhadap distribusi kueri.

Gu dkk. mempelajari kekuatan DNN dengan mempelajari strategi pra-pemrosesan dan pelatihan yang memperhitungkan struktur contoh adversarial serta topologi jaringan model yang ditargetkan. Pra-pemrosesan dilakukan dengan denoising autoencoders (DAEs). Autoencoder dipilih sebagai model pembelajaran mendalam karena mempertahankan distribusi data asli non-adversarial dengan memetakan kembali data pelatihan asli ke dirinya sendiri. Eksperimen di Gu et al. menunjukkan bahwa DAE mampu menghilangkan gangguan adversarial dalam strategi pelatihan DNN. Selain itu, prosedur pelatihan end-to-end dengan fungsi penalti yang menghaluskan data adversarial diusulkan dengan menumpuk DAE ke dalam jaringan saraf feedforward yang disebut autoencoder kontraktif (CAE). Hukuman tambahan di CAE meminimalkan norma kuadrat Jacobian dari representasi tersembunyi dari data masukan.

DNN mencapai kinerja tinggi karena rangkaian unit non-linier yang dalam memungkinkan generalisasi non-lokal dalam manifold khusus data. Kemampuan DNN untuk secara otomatis mempelajari prior generalisasi non-lokal dari data merupakan kekuatan dan kelemahan pembelajaran adversarial di lingkungan dunia nyata. Contoh adversarial di DNN disebabkan oleh alasan berikut oleh Gu et al.

- Dalam data berdimensi tinggi, asumsi kelancaran yang mendasari metode kernel tidak berlaku untuk arsitektur jaringan saraf feedforward deterministik
- Penerapan metode kernel pada ruang manifold tidak menjamin generalisasi lokal pada ruang masukan
- Karena sifat generalisasi lintas model dari contoh manipulasi pembelajaran , penyerang dapat menghasilkan contoh manipulasi pembelajaran dari model independen
- Lebih sedikit derajat kebebasan data yang ditangkap seiring dengan meningkatnya lapisan jaringan neural dalam

Oleh karena itu, menurut Gu dkk., tantangan dalam desain DNN adalah untuk melatih jaringan mendalam yang tidak hanya menggeneralisasi dalam ruang berjenis abstrak untuk mencapai akurasi pengenalan yang baik tetapi juga mempertahankan generalisasi lokal dalam ruang masukan. Baik dalam model dangkal maupun model dalam, contoh-contoh adversarial juga bersifat universal dan tidak dapat dihindari menurut definisi ini. Oleh karena itu, arsitektur pembelajaran mendalam yang kuat terhadap data adversarial harus dilatih untuk memasukkan invarian masukan sehubungan dengan keluaran jaringan akhir yang memperhitungkan data adversarial.

2.4 CONTOH MANIPULASI UNTUK PENGKLASIFIKASI YANG MENYESATKAN

Meskipun jaringan saraf mencapai kinerja tinggi dengan mengekspresikan komputasi sewenang-wenang dalam bentuk langkah-langkah non-linier paralel yang masif, Szegedy dkk. melakukan pengamatan bahwa lapisan jaringan saraf tidak menguraikan distribusi basis dari informasi semantik. Szegedy dkk. menemukan bahwa jaringan dalam mempelajari pemetaan input-output yang terputus-putus sehingga gangguan yang tidak terlihat meningkatkan kesalahan prediksi jaringan dalam bahkan ketika jaringan tersebut dilatih pada subset

kumpulan data yang berbeda. Gangguan yang tidak terlihat seperti itu disebut contoh adversarial.

Mempelajari contoh-contoh adversarial secara intrinsik terhubung dengan struktur jaringan dalam dan distribusi data masukan. Dalam percobaan oleh Szegedy et al., sejumlah besar contoh adversarial ditemukan salah klasifikasi oleh jaringan dalam. Contoh adversarial ini dibuat dengan mengubah pengaturan hyperparameter jaringan dalam seperti jumlah lapisan, inisialisasi bobot, dan regularisasi bobot. Jadi, Szegedy dkk. menyimpulkan bahwa contoh-contoh adversarial bukanlah hasil dari overfitting model pembelajaran mendalam tertentu.

Dalam sinyal masukan berdimensi tinggi ke model linier sederhana, Goodfellow et al. mengamati bahwa banyak perubahan yang sangat kecil pada masukan dari contoh-contoh adversarial menambah satu perubahan besar pada keluaran dalam pembelajaran mendalam. Teman baik dkk. berhipotesis bahwa pengklasifikasi jaringan dalam menunjukkan perilaku linier dalam ruang berdimensi tinggi. Contoh adversarial kemudian dianalisis sebagai properti perkalian titik berdimensi tinggi. Stabilitas bobot model yang mendasari dikatakan menghasilkan stabilitas contoh yang berlawanan. Untuk mendapatkan gangguan adversarial, fungsi biaya untuk melatih jaringan dalam dilinearisasikan di sekitar nilai parameter saat ini.

Metode untuk menghasilkan contoh adversarial ini disebut metode tanda gradien cepat (FGSM). Dalam FGSM, arah gangguan adversarial dihipotesiskan lebih penting daripada posisinya dalam ruang data. Kemudian, pelatihan model pembelajaran mendalam adversarial diusulkan sebagai regularisasi non-linier yang mengamankan jaringan dalam dengan meminimalkan kesalahan terburuk pada contoh adversarial FGSM. Pelatihan adversarial juga dipandang sebagai pembelajaran aktif di mana model pembelajaran memperoleh label baru untuk contoh adversarial dari pelabel heuristik yang menyalin label titik-titik terdekat. Papernot dkk. memperkenalkan strategi serangan kotak hitam untuk menghasilkan contoh adversarial tanpa sepengetahuan internal jaringan saraf dalam target.

Nguyen dkk. menghasilkan contoh-contoh adversarial dengan algoritma evolusioner dan menyebutnya sebagai “gambaran bodoh.” Gambar bodoh tidak dapat dikenali oleh mata manusia tetapi diklasifikasikan sebagai objek yang dapat dikenali dengan keyakinan tinggi oleh jaringan saraf dalam (DNN). Populasi gambar yang menipu berevolusi dengan merancang algoritma evolusi yang disebut arsip multidimensi elit fenotipik (MAP-Elite). MAP- Elit menyimpan individu terbaik yang ditemukan sejauh ini untuk setiap tujuan. Kemudian ia memutasi organisme yang dipilih secara acak dari populasi dan menggantikan jagoan saat ini untuk tujuan apa pun jika individu baru memiliki kebugaran yang lebih tinggi untuk tujuan tersebut. Skor prediksi DNN diambil sebagai fungsi kebugaran di MAP-Elite. Untuk kelas mana pun yang pernah dilihat sebelumnya, gambar bodoh yang dihasilkan dengan skor prediksi lebih tinggi akan menjadi juara untuk kelas tersebut.

Piksel gambar kumpulan data MNIST dan piksel gambar yang dihasilkan oleh jaringan penghasil pola komposisi (CPPN) mewakili genom di MAP-Elite. Berbagai fungsi aktivasi CPPN memberikan keteraturan geometris yang berbeda pada gambar yang menipu. Operator evolusi MAP-Elite menentukan topologi, bobot, dan unit aktivasi setiap jaringan CPPN dalam

populasi. Untuk berbagai hipotesis tentang hubungan antara kumpulan data pelatihan dan arsitektur DNN, skor prediksi dan pengujian Mann-Whitney U memvalidasi keluaran distribusi gambar yang dibodohi oleh MAP-Elite.

Carlini dkk. merancang skenario serangan kotak putih dan kotak hitam untuk jaringan saraf feedforward yang bertindak sebagai pengklasifikasi. Di berbagai mekanisme deteksi, fungsi kerugian manipulasi pembelajaran baru diusulkan untuk mengelabui pengklasifikasi jaringan saraf. Eksperimen kemudian diusulkan untuk mengeksplorasi ruang data dan sifat transferabilitas dari contoh-contoh adversarial yang ditargetkan. Untuk merumuskan serangan, ditentukan tiga model ancaman yaitu musuh tanpa pengetahuan, musuh dengan pengetahuan sempurna, dan musuh dengan pengetahuan terbatas. Musuh yang tidak memiliki pengetahuan tidak memiliki pengetahuan tentang keberadaan detektor saat menargetkan prediksi label kelas dari pengklasifikasi.

Oleh karena itu, musuh yang tidak memiliki pengetahuan bertindak sebagai garis dasar untuk menargetkan setiap detektor yang diusulkan. Sebagai perbandingan, musuh dengan pengetahuan sempurna memiliki pengetahuan penuh tentang parameter pengklasifikasi dan skema deteksi detektor. Musuh yang berpengetahuan sempurna kemudian melakukan serangan kotak putih. Untuk melakukan serangan kotak hitam, Carlini dkk. berasumsi bahwa musuh dengan pengetahuan terbatas mengetahui skema deteksi detektor tetapi tidak memiliki akses ke pengklasifikasi terlatih, detektor terlatih, atau data pelatihan mereka.

Skema detektor dipelajari oleh Carlini et al. mencakup (i) jaringan saraf kedua untuk mengklasifikasikan gambar sebagai alami atau adversarial, (ii) analisis komponen utama (PCA) untuk mendeteksi sifat statistik gambar atau parameter jaringan, (iii) uji hipotesis statistik (seperti uji perbedaan rata-rata maksimum dan model campuran Gaussian yang membandingkan distribusi data manipulasi pembelajaran dengan distribusi data asli), dan (iv) normalisasi masukan dengan pengacakan dan pengaburan. Jarak antara contoh manipulasi pembelajaran dan contoh pelatihan diasumsikan sebagai fungsi kerugian manipulasi pembelajaran yang mengukur ketahanan pertahanan dalam mekanisme deteksi.

Dalam serangan penurunan gradien berulang, fungsi kerugian manipulasi pembelajaran memiliki istilah regularisasi tambahan yang membandingkan kemungkinan log jaringan dalam dalam memprediksi kelas target dengan kelas yang paling mungkin berikutnya. Ambang batas yang ditentukan pengguna pada kemungkinan log peringkat kemudian menetapkan tingkat keyakinan tinggi atau rendah pada contoh adversarial yang dihasilkan. Evaluasi sifat-sifat contoh manipulasi pembelajaran direkomendasikan untuk dilakukan sesuai dengan kriteria evaluasi berikut:

- Mengevaluasi contoh-contoh adversarial di beberapa kumpulan data (seperti MNIST dan CIFAR) dengan pertahanan yang tidak beroperasi secara langsung pada piksel
- Mengevaluasi skema baru untuk kekuatan serangan yang menunjukkan musuh mampu melakukan serangan untuk menghindari deteksi ketika menyadari adanya pertahanan yang diusulkan
- Melaporkan angka positif palsu selain angka positif sebenarnya dalam evaluasi kinerja.

Di sini, jaringan saraf dikatakan kuat jika menemukan contoh adversarial yang melewati detektornya merupakan proposisi yang sulit.

Baluja dkk. mengusulkan serangan yang ditargetkan di mana jaringan saraf feedforward yang disebut jaringan transformasi adversarial (ATN) dilatih untuk menghasilkan contoh adversarial. ATN menghasilkan contoh adversarial yang sedikit mengubah keluaran pengklasifikasi berdasarkan masukan asli. Sebaliknya, Moosav dkk. membuat teknik serangan tidak bertarget, yaitu DeepFool, yang dioptimalkan dengan metrik jarak antara contoh adversarial dan contoh normal.

Dalam penelitian kami di, kami menghasilkan contoh adversarial untuk mempengaruhi serangan keracunan pada data pelatihan klasifikasi. Contoh-contoh adversarial dihasilkan oleh manipulasi adversarial yang dipelajari selama serangan teoretis permainan kami terhadap proses pelatihan pelajar. Dalam skenario serangan kotak hitam yang menghasilkan distribusi data pengujian, diasumsikan tidak ada pengetahuan sebelumnya tentang model pembelajaran. Musuh kita tidak mengetahui proses pelatihan model pembelajaran maupun strategi respons terbaik model pembelajaran di seluruh permainan Stackelberg.

Musuh kita melakukan serangan yang ditargetkan untuk memanipulasi beberapa label positif menjadi satu label negatif. Kekuatan serangan manipulasi adversarial kami ditentukan dalam parameter pengacakan pencarian di ALS dan SA. Skalar optima di SA digunakan untuk menghasilkan vektor optima di ALS. Optima lokal di ALS menyatu dengan Optima stokastik non-cembung yang menyelesaikan permainan Stackelberg untuk menghasilkan keluaran manipulasi adversarial yang optimal. Manipulasi adversarial yang optimal mampu mengkodekan data adversarial dalam kaitannya dengan parameter statistik multivariat dari model campuran Gaussian yang dihasilkan dalam kumpulan data multi-label.

2.5 JARINGAN MANIPULASI GENERATIF

Teman baik dkk. menyatakan bahwa penyebab utama kerentanan jaringan pembelajaran mendalam terhadap contoh-contoh adversarial adalah sifat liniernya dalam ruang pencarian berdimensi tinggi. Jaringan pembelajaran mendalam juga berkinerja buruk pada contoh data pengujian yang tidak memiliki probabilitas tinggi dalam distribusi data pelatihan. Oleh karena itu, contoh adversarial dapat dihasilkan dengan menerapkan gangguan kasus terburuk pada data pelatihan. Masukan yang terganggu menghasilkan prediksi keluaran yang salah dengan keyakinan tinggi. Jadi, Goodfellow dkk. berpendapat perlunya memiliki prosedur pelatihan adversarial yang tujuannya adalah untuk meminimalkan kesalahan terburuk ketika data pelatihan diganggu oleh musuh. Teman baik dkk. kemudian merumuskan pelatihan adversarial sebagai permainan min-maks antara dua jaringan saraf dalam. Model generatif mendalam yang dihasilkan disebut generative adversarial network (GANs).

Berbagai metode generatif mendalam tersedia untuk menciptakan gangguan antara distribusi data pelatihan dan pengujian. Radford dkk. mengusulkan GAN stabil yang disebut DCGAN. Gulrajani dkk. merancang IWGAN yang melakukan analisis teoritis tentang proses

pembelajaran generatif. Berthelot dkk. mengusulkan MULAI dengan fungsi kerugian baru dalam algoritma pelatihan. Chen dkk. mengusulkan InfoGAN yang menggunakan model pembelajaran generatif untuk pembelajaran representasi tanpa pengawasan.

Sejauh menyangkut mekanisme pertahanan pembelajar, formulasi game kami mirip dengan formulasi game GAN. Namun, tujuan penelitian kami adalah untuk mensimulasikan skenario serangan adversarial nyata pada model klasifikasi dua label dan multi-label dalam hal kerugian yang ditanggung musuh. Kami berupaya meningkatkan kinerja klasifikasi ketika distribusi data diubah dengan niat jahat. Sebaliknya, tujuan GAN adalah menghasilkan data sintetik yang tidak dapat dibedakan dari data asli. Fungsi tujuan kami memiliki istilah biaya dan kesalahan yang menentukan skenario serangan dalam pengaturan pembuatan data yang berlawanan. Sebaliknya, fungsi tujuan di GAN didefinisikan dalam bentuk fungsi kerugian jaringan saraf dalam yang mempelajari distribusi data pelatihan dan pengujian tertentu.

Dalam formulasi permainan min-maks, kami berupaya membuat kumpulan data untuk skenario serangan dalam model pembelajaran diskriminatif dan masalah pembelajaran terawasi, sementara GAN menangani model pembelajaran generatif dan masalah pembelajaran tanpa pengawasan. Selain itu, generator adalah pemimpin permainan dalam formulasi min-max untuk GAN, sedangkan dalam formulasi min-max kami, musuh yang cerdas memimpin permainan. Saat mencari keseimbangan Nash dalam permainan min-maks, GAN memecahkan masalah optimasi cembung dengan algoritma optimasi berbasis gradien, sedangkan kami memecahkan masalah optimasi stokastik non-cembung dengan algoritma pembelajaran evolusi. Dengan demikian, kita dapat memperkirakan biaya terbaik bagi musuh dalam melakukan serangan musuh.

Jaringan adversarial generatif (GAN) memperkirakan kemungkinan data dengan kerangka adversarial yang melibatkan permainan dua pemain antara jaringan generator G dan jaringan diskriminator D. IWGAN meningkatkan estimasi GAN dengan regularisasi yang tidak menimbulkan korelasi antara contoh yang dihasilkan. Tujuan dari formulasi permainan kami dengan autoencoder variasional bukanlah untuk meningkatkan akurasi klasifikasi dengan menambah data asli yang melatih autoencoder. Yang penting, kami mencatat bahwa perbedaan mendasar antara penelitian kami dengan autoencoder variasional dan tujuan jaringan generatif adalah menipu pengklasifikasi daripada meniru data asli. Kami memecahkan masalah pembelajaran yang diawasi dengan lawan yang bervariasi, sedangkan model generatif mendalam umumnya menyelesaikan masalah pembelajaran tanpa pengawasan atau pembelajaran semi-supervisi dengan musuh generatif.

2.6 JARINGAN MANIPULASI PEMBELAJARAN MESIN GENERATIF

Contoh adversarial telah didefinisikan untuk model generatif mendalam. Distribusi manipulasi adversarial dalam serangan kotak putih dan juga serangan kotak hitam telah dimodelkan dengan AdvGAN. Rangkaian penelitian tentang autoencoder adversarial menerapkan distribusi sebelumnya pada keluaran data pelatihan pembelajaran jaringan encoder, di mana autoencoder secara diskriminatif memprediksi apakah sampel berasal dari ruang latennya atau dari distribusi sebelumnya yang ditentukan oleh pengguna. Sebaliknya,

masalah pengoptimalan teoretis permainan kami tidak bergantung pada distribusi data pelatihan dan model klasifikasi tertentu.

Larsen dkk. mengusulkan pembelajaran adversarial generatif dalam rekonstruksi hilangnya autoencoder variasional. Tran dkk. mengusulkan batasan pada fungsi jarak untuk melatih jaringan adversarial generatif di ruang laten autoencoder. Gregor dkk. mengusulkan autoencoder berbasis mekanisme perhatian untuk mempelajari ruang laten dalam kerangka autoencoder variasional berurutan. Ha dkk. mengusulkan jaringan saraf berulang untuk pembuatan sketsa dalam gambar. Makhzani dkk. mengusulkan mekanisme pelatihan adversarial untuk autoencoder probabilistik.

Taksonomi skenario serangan adversarial dalam pembelajaran mendalam disediakan oleh Gilmer et al. dan Biggio dkk. Skenario serangan kami dengan permainan Stackelberg mengusulkan fungsi pembayaran adversarial yang baru. Kami mewakili ruang fitur untuk manipulasi adversarial dalam hal fungsi biaya adversarial, operator stokastik, dan strategi permainan dalam algoritma simulasi anil.

Wang dkk. teori survei dan implementasi jaringan adversarial generatif. Taksonomi formulasi jaringan adversarial generatif yang relevan untuk algoritma pembelajaran adversarial teoretis permainan dirangkum dalam Tabel 2.1, dan 2.2. Di seluruh baris tabel, perbandingan algoritme dibuat berdasarkan skenario serangan jaringan generator, fungsi kerugian, ruang strategi, dan fungsi tujuan. Sebagian besar model generatif mendalam tidak menganalisis distribusi data dalam kaitannya dengan optimalisasi teoretis permainan dari fungsi tujuan. Sebagai perbandingan, metode kami mengusulkan fungsi pembayaran adversarial untuk optimasi dan fungsi biaya adversarial untuk regularisasi dalam fungsi tujuan.

Pembelajaran Fitur Kausal dan Pembelajaran Mesin

Metode kausalitas telah diterapkan pada masalah pembelajaran mendalam seperti pembelajaran semi-supervisi dan pembelajaran transfer. Dalam masalah ini, informasi sebelumnya yang diambil dari jaringan lain digunakan untuk memusatkan bobot dalam jaringan pembelajaran mendalam hibrid. Jaringan tersebut kemudian digunakan untuk membangun hipotesis statistik tentang pola, struktur, konteks, dan konten dalam data aktual.

Algoritme pembelajaran propagasi mundur untuk jaringan dalam telah ditingkatkan dengan melatih model grafis probabilistik. Pelatihan semacam itu pada dasarnya adalah Bayesian di mana distribusi sebelumnya menginformasikan dan membatasi model analitik yang memprediksi distribusi posterior. Algoritme pembelajaran mendalam yang ditingkatkan menghasilkan keluaran yang diprediksi berdasarkan inferensi kausal. Dalam kerangka Bayesian, metode kausalitas juga meningkatkan interpretasi jaringan dalam yang beroperasi di lingkungan yang tidak pasti.

Tabel 2.1 Perbandingan jaringan adversarial generatif

Jaringan Adversarial	Skenario serangan	Fungsi kerugian	Fungsi kehilangan generator	Fungsi divergensi informasi	Tipe permainan	Fungsi pembayaran	Fungsi biaya	Kendala optimasi

Pertahanan GAN (536)	Modelkan distribusi gambar yang baik, terganggu dalam serangan kotak putih dan serangan kotak hitam	Sama seperti WGAN	Sama seperti WGAN	MSE, L2 norma	Permainan min-maks	Jaringan reformis, kode laten	Pelatihan adversarial menambah data pelatihan	Perwakilan GAN untuk merekonstruksi contoh-contoh adversarial
GENG (474	Respons terbaik yang dibatasi sumber daya pada data sintetis	Skor pengklasifikasi itulah fungsi dari kedua data nyata dan data palsu	Pembayaran generator sebagai fungsi dari data palsu saja	Ekuilibrium Nash yang dibatasi sumber daya deterministik dan nondeterministik	Permainan strategi zero-sum	Imbalan pemain dalam profil strategi campuran, Definisi 2	Fungsi pengukuran pada Definisi 6 dan Teorema 10	GANG terbatas pada data diskrit
AdvGAN (659)	Gangguan adversarial semi-kotak putih dengan batas target dan kebenaran dasar	Pelatihan model distilasi statis dan dinamis dengan pendekatan minimalisasi alternatif	Kerugian serangan yang ditargetkan untuk LSGAN diberikan pada Persamaan 4	Menyatukan pembelajaran adversarial	Permainan min-maks	Sama seperti Goodfellow GAN	Kehilangan engsel	Kerugian lintas entropi
DeLIGAN (245)	Data pelatihan yang terbatas untuk menangkap keragaman di seluruh modalitas gambar	Sama seperti DCGAN	Sama seperti DCGAN	m-IS merupakan divergensi KL yang mengukur intra-kelas sample diversity along with the sample quality	Permainan min-maks	Parameterisasi ulang ruang laten pada distribusi sebelumnya	Regulator L2 untuk mencegah maxima lokal pada generator	Bobot campuran seragam dalam penurunan gradien
EBGAN [702]	Pembuatan sampel kontrastif yang dibuat dengan tangan dan diatur dalam pengaturan yang diawasi, diawasi dengan lemah, dan tanpa pengawasan	Rekonstruksi fungsi kehilangan dan energi dalam autoencoder	Sama seperti Goodfellow GAN	Skor awal	Permainan min-maks	Pelatihan adversarial	Kesalahan rekonstruksi	Pencarian grid pada pilihan arsitektur dan hyperparameter
Fisher GAN [445]	Perbedaan standar dalam pengujian hipotesis dua sampel dan pembelajaran semi-supervisi	Jarak Mahalanobis antar fitur berarti penyematan distribusi asli dan palsu	Berarti jarak penyematan	Metrik probabilitas integral (IPM) dan skor awal yang diparametrikan oleh jaringan saraf generatif	Permainan min-maks	Sama seperti DCGAN	Sama seperti DCGAN	Momen kritik urutan kedua yang membedakan dua distribusi, istilah regularisasi lintas entropi dalam pembelajaran semi-supervisi
Peningkatan f-GAN [496]	Divergensi generator diurutkan dari pencarian mode terbanyak hingga cakupan mode terbanyak	Sama seperti Goodfellow GAN	Perkiraan rasio kepadatan model terhadap kepadatan data dari diskriminator saat ini	f-Divergensi	Permainan min-maks	Maksimalisasi harapan	Kemungkinan pembelajaran data palsu yang nyata sangat sulit dilakukan	Metrik/faktor kualitas/keanekaragaman gambar tanpa mode pelepasan

Kami tertarik pada skenario serangan dengan model variabel laten dalam pembelajaran adversarial teoritis permainan. Kumari dkk. mempelajari serangan kotak putih pada tingkat lapisan laten model klasifikasi gambar yang dilatih secara musuh. Ketahanan yang lebih tinggi pada lapisan fitur dicapai dengan pelatihan adversarial lapisan laten dengan varian FGSM yang berulang. Sebaliknya, penelitian kami menciptakan model generatif yang mendalam untuk manipulasi adversarial yang menyediakan pengatur teoretis permainan pada fungsi kerugian pengklasifikasi yang ditargetkan.

Chattopadhyay dkk. mengusulkan model sebab akibat struktural untuk pengaruh sebab akibat dari fitur masukan pada keluaran jaringan saraf. Pengaruh kausal pada keluaran fungsi prediksi disebut atribusi jaringan saraf. Mereka dikatakan sebagai artefak penyebab jaringan dalam yang lebih dapat diinterpretasikan daripada fitur regresi yang terutama memetakan korelasi antara masukan dan keluaran jaringan saraf. Dalam tugas prediksi urutan dengan model kausal struktural seperti itu, ketergantungan kausal antara neuron masukan yang berbeda diasumsikan secara bersama-sama disebabkan oleh perancu laten seperti mekanisme penghasil data yang diterapkan pada model deret waktu.

Yang dkk. mempelajari fitur tingkat piksel untuk penalaran kausal dalam penyembunyian piksel dan gangguan adversarial. Ancona dkk. dan Lundberg dkk. membahas metode atribusi dalam nilai-nilai Shapley dari teori permainan kooperatif. Investigasi penelitian kami adalah menciptakan artefak manipulasi adversarial teoritis permainan yang dapat ditafsirkan. Untuk mencapai tujuan ini, kami telah membuat fitur kausal Granger dari prediksi regresi. Di masa depan, kami akan membuat garis dasar prediktif dalam model variabel laten dari mekanisme penghasil data dalam atribusi jaringan saraf. Kami berharap data dasar tersebut akan menemukan fitur kontrafaktual dalam pengklasifikasi berbasis aturan khusus aplikasi.

Tabel 2.2 Perbandingan jaringan adversarial generatif

Adversarial Network	Skenario Serangan	Fungsi kerugian diskriminator	Fungsi Generator loss	Fungsi Divergensi Informasi	Tipe Permainan	Fungsi Pembayaran	Fungsi Biaya	Kendala Optimasi
CausalGAN	Grafik sebab Akibatyang benar/layak yang terstruktur pada label data distribusi observasi dan intervensi pada gambar dan tabel secara bersamaan	Kerugian GAN bersyarat dicampur dengan kehilangan label	Sama seperti DCGAN dengan gradien label	Total Variasi Jarak (TVD)	Sama seperti WGAN, DCGAN	Sama seperti WGAN, DCGAN	Prosedur dua tahap pada distribusi gambar terkondisi label ganda dan label biner	Margin tupel koefisien dalam penurunan gradien stokastik
AM-GAM	Pelabelan yang telah ditentukan sebelumnya dan pelabelan dinamis untuk pembuatan gambar, kualitas gambar dan keragaman gambar	Entropi silang untuk klasifikasi kelas jamak	Sama seperti DCGAN	Skor awal, skor AM	Permainan Min Max	Tidak ada	Tidak ada	Gradien standart kelas

MBGAN	Tidak ada	Metrik nonlinier pada persamaan 9 dan 10	Sama seperti DCGAN	Skor Awal	Permainan Min Max	Tidak ada	Tidak ada	Weight Cutting, Penalty Center
Triangle-GAN	Klasifikasi Gambar semi supervisi, terje,aham gambar ke gambar dan pembuatan gambar berbasis atribut	Kerugiana GAN bersyarat dan GAN dua arah	Dua generator GAN bersyarat yang berhubungan dengan dua diskriminator	Divergensi jensen-shannon (JSD) ditambah divergensi Kullback Leibler (KL)	Permainan Min Max	Tidak ada	Tidak ada	Tidak ada
f-CLSWGAN	Tidak ada contoh berlabel kelas tertentu dalam penyematan multimodal	Pengklasifikasian softmax untuk pembelajaran zero-shot	Sama seperti WGAN	Model penyematan multimodal dengan contoh berlabel dari kelas yang dilihat dan fitur CNN mendalam yang dikondisikan pada informasi semantik tingkat kelas	Permainan Min Max	Tidak ada	Data yang dihasilkan memiliki dimensi yang jauh lebih rendah dibandingkan gambar berkualitas tinggi yang diperlukan untuk diskriminasi	Penyematan kelas yang memodelkan hubungan semantik antar kelas
LSGAN	Menghukum sampel berdasarkan jaraknya ke batas keputusan	Persamaan 1.2	Persamaan 1.2 untuk menghasilkan sampel menuju batas keputusan dan keragaman data nyata	f-Divergensi, divergensi chi-kuadrat Pearson	Permainan Min Max	Tidak ada	Tidak Ada	Persamaan deterministik antara label untuk data palsu, data nyata, dan data yang dihasilkan untuk pengkodean one-hot dan reduksi dimensi
D2GAN	Pembuatan gambar	KullbackLeibler (KL) dan membalikkan divergensi KL	Fungsi kepadatan multi-mode apa pun	Skor awal dan skor MODE	Permainan Min Max	Tidak Ada	Tidak ada	Bola penutup minimal, sasaran pengganti, banyak pemain, dll. dalam fungsi kepadatan generator
GAN	Kerangka kerja adversarial untuk estimasi kemungkinan yang menyebarkan turunan diskriminator melalui proses generatif	Jaringan aktivasi maxout	Aktivasi linier penyearah dan jaringan aktivasi sigmoid	Divergensi KullbackLeibler, divergensi JensenShannon	Permainan min max	Tidak ada	Tidak ada	G dan D diberikan kapasitas dan waktu pelatihan yang cukup; tidak ada overfitting di D; G tidak boleh dilatih terlalu banyak tanpa memperbarui
IWGAN	Algoritme pelatihan GAN yang lebih baik	Kekurangan Kritis di WGAN	Kerugan adversarial di WGAN	Jarak penggerak bumi	Permainan min max	Tidak ada	Tidak ada	Penalti pada norma gradien, skema normalisasi yang tidak menimbulkan korelasi antar contoh

InfoGAN	Pisahkan representasi yang dapat ditafsirkan dari data yang tidak berlabel	Sama Seperti DCGAN	Generator informasi yang diatur tentang kebisingan yang tidak dapat dimampatkan dan fitur semantik terstruktur	Informasi timbal balik dan entropi diferensial	Min Max Permainan	Tidak ada	Tidak Ada	Algoritma sleep-awake dengan regularisasi variasional
ss-InfoGAN	Representasi data yang diekstraksi dan dikontrol di mana variabel laten sesuai dengan kategori label dan mempelajari kode kontinu dan kategoris	Sama dengan DCGAN	Informasi timbal balik antara vektor kode dan sampel berlabel nyata dan sampel tak berlabel sintetik, entropi silang untuk kode laten kategorikal, kesalahan kuadrat rata-rata untuk kode laten berkelanjutan	Informasi timbal balik dan entropi diferensial	Permainan Min Max	Tidak ada	Tidak ada	Faktor-faktor yang tidak sesuai dengan label tidak akan ditemukan dalam pengaturan yang diawasi dan semi-diawasi, distribusi data nyata dan sintetis bersifat independen, dan label mengikuti distribusi tetap sehingga memiliki entropi tetap
MCGAN	Statistik fitur mean dan kovarians	Sama seperti WGAN dengan hilangnya entropi silang pada data berlabel	Norma rata-rata Lq, norma kovarians Ky-Fan (norma inti dari perbedaan kovarians terpotong), fitur pencocokan metrik probabilitas integral (IPM); IPM adalah fungsi linier terbatas yang ditentukan dalam ruang fitur nonlinier yang disebabkan oleh peta fitur parametrik	Jarak geodesik antara kovarians dan ukuran probabilitas dalam pengaturan multimodal	Permainan Min Max	Tidak ada	Tidak Ada	Mode terbatas dari penyematan fitur pada distribusi nyata dan palsu, sampel yang memadai dari data "nyata" dan "palsu" untuk melatih ruang fitur "generator" dan "kritikus"
DCGAN	Representasi fitur yang dapat digunakan kembali dari kumpulan data besar yang tidak berlabel, pengelompokan hierarki dari	Aktivasi yang diperbaiki dan bocor	Dekonvolusi dan pemfilteran aktivasi maksimal dari setiap filter konvolusi di jaringan	Persentase akurasi pada data pelatihan, pengujian, validasi	Permainan Min Max	Tidak Ada	Tidak ada	Batch normalization

	representasi perantara							
BEGAN	Mencocokkan distribusi kesalahan dan bukan distribusi sampel, menimbang secara dinamis istilah regularisasi atau tujuan heterogen lainnya	Diskriminator memiliki dua tujuan yang bersaing dalam kontrol umpan balik loop tertutup: mengkodekan gambar nyata secara otomatis dan membedakan gambar nyata dari gambar yang dihasilkan	Negatif dari kerugian diskriminator	Pengukuran konvergensi global dengan menggunakan konsep keseimbangan batas dari teori kontrol proporsional	Permainan min max	Tidak ada	Tidak ada	Pemilihan hyperparameter yang benar untuk menjaga keseimbangan antara kerugian generator dan diskriminator
BGAN	Sama seperti DCGAN	Sama seperti DCGAN	Tujuan PERKUAT yang mencari batas dengan pelatihan gradien kebijakan di mana imbalan adalah bobot kepentingan yang dinormalisasi	f-Divergence dan divergensi Jensen-Shannon dengan bobot kepentingan untuk sampel yang dihasilkan	Permainan MinMax	Tidak ada	Tidak ada	Memperkirakan ekspektasi dalam bobot kepentingan yang dinormalisasi dengan menggunakan sampel Monte-Carlo
AutoGAN	Penanda pencitraan yang relevan untuk perkembangan penyakit dan pemantauan pengobatan	Hilangnya diskriminasi memaksa citra yang dihasilkan terletak pada keragaman yang dipelajari	Kehilangan sisa yang memaksakan kesamaan visual antara gambar yang dihasilkan dan gambar kueri	Skor deteksi anomali/kebaruan dari fungsi klasifikasi pencocokan fitur di ruang laten	Sama seperti DCGAN	Tidak ada	Tidak ada	Representasi fitur yang sesuai dalam diskriminator

Kecerdasan Buatan dan Manipulasi Pembelajaran Mesin yang Dapat Dijelaskan

Kami tertarik pada kecerdasan buatan yang dapat dijelaskan (XAI) dari model generatif mendalam yang dapat diterapkan pada pembelajaran adversarial teoretis dalam serangan kotak hitam. Lou dkk. memperkenalkan model aditif umum (GAM) sebagai perpanjangan model linier umum (GLM) yang dapat ditafsirkan. Guidotti dkk. mensurvei kemampuan menjelaskan model kotak hitam. Rudin membandingkan model XAI dengan model yang dapat diinterpretasikan secara inheren. Wang dkk. mengusulkan kumpulan aturan hibrid yang mengintegrasikan model yang dapat ditafsirkan dengan model kotak hitam. Frost dkk. membuat pohon keputusan yang menggeneralisasi pembelajaran jaringan saraf. Ribeiro dkk. memberikan penjelasan tekstual untuk klasifikasi gambar dan menjawab pertanyaan visual. Ignatiev dkk. mengusulkan sistem penalaran kendala untuk menjelaskan prediksi.

Strumbelj dkk. menjelaskan prediksi dengan teori permainan koalisi. Bulo dkk. mendefinisikan permainan prediksi acak yang merupakan formulasi teoretis permainan non-kooperatif di mana pengklasifikasi dan penyerang membuat pilihan strategi acak berdasarkan beberapa distribusi probabilitas yang ditentukan pada strategi masing-masing yang diatur

dalam pengenalan digit tulisan tangan, deteksi spam, dan deteksi malware. Peake dkk. membuat struktur aturan asosiasi yang dapat ditafsirkan dari sistem rekomendasi faktor laten yang melatih model kotak hitam faktorisasi matriks. Lakkaraju dkk. membuat model berbasis aturan dengan pembelajaran kumpulan keputusan yang dirancang untuk interpretasi optimasi fungsi submodular. Baehren dkk. mengusulkan metode penjelasan untuk keputusan metode klasifikasi apa pun. Ribeiro dkk. menjelaskan prediksi pengklasifikasi apa pun sebagai masalah optimasi submodular. Shrikumar dkk. menghitung skor penting untuk aktivasi neuron di jaringan saraf yang menunjukkan keunggulan signifikan dibandingkan metode berbasis gradien. Koh dkk. menggunakan fungsi pengaruh dari statistik yang kuat untuk menjelaskan prediksi.

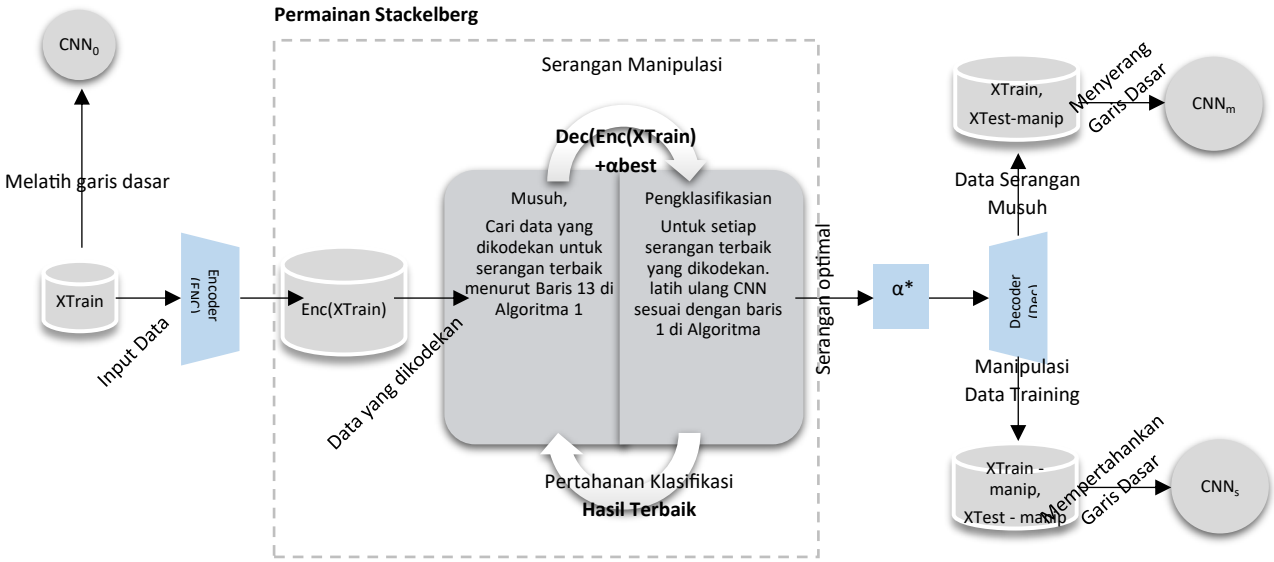
Bastani dkk. mengusulkan metrik untuk mengevaluasi ketahanan jaringan saraf dalam. Narodytska dkk. membuat representasi Boolean dari jaringan saraf dalam untuk memverifikasi propertinya. Tomsett dkk. mensurvei hubungan antara interpretabilitas dan serangan adversarial. Liu dkk. mengembangkan kerangka deteksi ketahanan musuh dengan memanfaatkan interpretasi model pembelajaran mesin. Tao dkk. mengusulkan teknik deteksi sampel adversarial untuk model pengenalan wajah, berdasarkan interpretabilitas. Fidel dkk. mengusulkan metode untuk mendeteksi contoh adversarial dengan nilai SHapley Additive exPlanations (SHAP) yang dihitung untuk lapisan internal DNN. Ilyas dkk. menghubungkan contoh-contoh adversarial dengan kehadiran fitur-fitur yang tidak kuat. Ignatiev dkk. menunjukkan bahwa penjelasan (XP) dari prediksi model pembelajaran mesin (ML) dan contoh adversarial (AE) dihubungkan oleh kerangka kerja logika orde pertama (FOL) yang disebut dualitas himpunan hit.

Ilustrasi Game Stackelberg dalam Pembelajaran Mesin Deep Learning

Gambar 2.3 adalah diagram alur untuk model permainan Stackelberg berbasis autoencoder adversarial kami. Pengklasifikasi multi-label CNNoriginal (selanjutnya disingkat CNNo) dengan bobot w∗ W dilatih pada data pelatihan berlabel Xtrain dan dievaluasi pada data pengujian berlabel Xtest yang bersumber dari database gambar. CNNo berpartisipasi dalam permainan dua pemain dengan musuh teoretis permainan kami. Musuh menyerang CNNo pada pos target label positif yang ditargetkan dengan menghasilkan serangan optimal α∗ A pada Nash Gambar 2.3 adalah diagram alur untuk model permainan Stackelberg berbasis autoencoder adversarial kami. Pengklasifikasi multi-label CNNoriginal (selanjutnya disingkat CNNo) dengan bobot w∗ W dilatih pada data pelatihan berlabel Xtrain dan dievaluasi pada data pengujian berlabel Xtest yang bersumber dari database gambar. CNNo berpartisipasi dalam permainan dua pemain dengan musuh teoritis permainan kami. Musuh menyerang CNNo pada pos target label positif yang ditargetkan dengan menghasilkan serangan optimal α∗ A pada data musuh Nash Enc(Xtrain) αterbaik dalam ruang yang dikodekan. Ini kemudian diterjemahkan sebagai Dec(Enc(Xtrain) αbest) untuk dievaluasi terhadap CNNo.

Setelah keluaran permainan konvergensi, α∗ optimal disimpulkan untuk setiap pasangan pos dan neg. Semua α∗ kemudian digabungkan untuk melakukan serangan adversarial multi-label pada CNNo untuk menghasilkan pengklasifikasi yang dimanipulasi CNNmanipulated (selanjutnya disingkat menjadi CNNm). CNNm akhirnya dilatih ulang

menjadi pengklasifikasi aman CNNsecure (selanjutnya disingkat CNN) yang tahan terhadap serangan adversarial multi-label.



Gambar 2.3 Diagram alur yang mengilustrasikan pemodelan teoretis permainan Stackelberg berbasis autoencoder adversarial

2.7 PEMBELAJARAN TRANSFER UNTUK ADAPTASI DOMAIN

Dalam pembelajaran mesin, pembelajaran transfer menerapkan pengetahuan yang dapat dipelajari yang diperoleh dari satu masalah analisis data ke masalah lain. Menyimpan, menggunakan kembali, dan mentransfer informasi dan pengetahuan dari kumpulan data dan tugas sebelumnya berpotensi meningkatkan efisiensi sampel dalam masalah pembelajaran mesin baru seperti yang melibatkan agen pembelajaran penguatan. Setelah pembelajaran yang diawasi, pembelajaran transfer merupakan pendorong besar keberhasilan dalam pembelajaran mesin komersial dan pembelajaran mendalam yang dapat diskalakan. Sebagai bentuk pembelajaran multi-tugas, pembelajaran transfer dapat digunakan dalam pembelajaran terawasi untuk meningkatkan klasifikasi multi-label dalam aplikasi pembelajaran mesin manipulatif seperti pemfilteran spam dan pengklasifikasi multikriteria.

Adaptasi domain adalah bidang pembelajaran transfer yang dapat diterapkan pada pemfilteran spam. Di dalamnya, distribusi sumber digunakan untuk mempelajari model yang berkinerja baik untuk distribusi target yang terkait tetapi berbeda dari distribusi sumber. Distribusi sumber dapat berupa email spam yang diterima oleh pengguna sumber di mana adaptasi domain berupaya memodelkan email spam untuk pengguna target yang berbeda. Dengan demikian, distribusi data sumber dan target memiliki ruang fitur yang sama tetapi distribusi datanya berbeda dalam adaptasi domain. Berbeda dengan adaptasi domain, ruang fitur data sumber untuk pembelajaran transfer bisa sama atau berbeda dengan data target.

Adaptasi domain dapat digunakan untuk memodelkan pergeseran distribusi dalam data yang tersedia untuk pelatihan dan algoritme pembelajaran mesin serta memvalidasi ketahanan distribusi yang sesuai dari algoritme pembelajaran adversarial. Jadi komunitas

pembelajaran mesin modern memiliki beberapa strategi untuk mendapatkan adaptasi domain antara kumpulan data pelatihan dan kumpulan data validasi dalam aplikasi praktis kecerdasan buatan. Strategi seperti itu membawa kita pada varian algoritma pembelajaran yang diawasi yang bersyarat, semi-terawasi, diawasi dengan lemah, multimodal, dan multi-terstruktur. Ini adalah bentuk pembelajaran terawasi yang lebih lemah di mana data pelatihan yang diberi label tangan tidak tersedia tanpa kesalahan dan gangguan pada label kelas.

Hal ini menghasilkan paradigma pembelajaran mesin yang bermusuhan seperti pembelajaran tambahan, pembelajaran utilitas, pembelajaran penguatan, dan pembelajaran online dengan informasi distribusi kelas dan biaya untuk representasi fitur yang dapat ditransfer dalam contoh adversarial karena deteksi outlier, deteksi kebaruan, dan deteksi titik perubahan dalam pergeseran distribusi. penyaringan informasi. Di sini, kita dapat memasukkan keahlian domain dengan fungsi untuk memberi label pada data pelatihan baru yang dihasilkan.

Contoh manipulasi pembelajaran mesin dalam pembelajaran Transfer

Tramer dkk. mengusulkan metode untuk menemukan dimensi contoh adversarial yang dapat ditransfer antar model pembelajaran mendalam. Subruang adversarial dengan jumlah dimensi yang besar lebih cenderung memungkinkan transferabilitas antar model pembelajaran mendalam. Analisis batas keputusan pembelajaran yang diawasi digunakan untuk mempelajari batas kemampuan transfer antar distribusi data. Model pembelajaran mendalam dalam visi komputer digunakan untuk membuat contoh-contoh adversarial yang dapat dikenali manusia tetapi komputer salah mengklasifikasikannya. Agen pembelajaran penguatan yang beroperasi di lingkungan data teoretis permainan diusulkan untuk membuat komputer salah mengklasifikasikan contoh-contoh adversarial. Contoh adversarial ditemukan terjadi di wilayah subruang fitur yang berdekatan dan relevan untuk mentransfer pembelajaran di antara titik-titik yang salah diklasifikasikan.

Subruang ini ditemukan untuk model linier dan kuadrat untuk pembelajaran mendalam pada masalah klasifikasi digit menggunakan dataset MNIST dan masalah deteksi malware menggunakan dataset DREBIN. Gangguan adversarial model-agnostik diperoleh dengan menggeser titik data pelatihan ke arah yang diperoleh dari perbedaan rata-rata kelas berlabel di ruang fitur masukan. Berbagai arah independen untuk menyusun manipulasi data adversarial juga diperoleh untuk mengukur dimensi subruang adversarial. Mereka dihasilkan dengan varian metode tanda gradien cepat yang membatasi gangguan adversarial dengan norma lp dari fungsi kerugian klasifikasi. Keteralihan contoh-contoh adversarial antara model pembelajaran mendalam dipelajari dengan jarak yang diusulkan antara batas-batas keputusan model yang tidak dipertahankan dan model yang dilatih secara adversarial.

Ma dkk. menilai dimensi contoh adversarial dengan distribusi jarak contoh adversarial dan tetangganya. Batas-batas keputusan dari subruang adversarial ternyata dapat dialihkan tergantung pada kedekatan titik data yang sah dengan titik-titik tersebut dalam arah adversarial. Keteralihan tersebut meningkat seiring dengan jumlah arah adversarial ortogonal independen dari subruang ini. Wang dkk. membuat contoh adversarial untuk model pembelajaran transfer yang digunakan dalam konteks aplikasi pengenalan gambar seperti

pengenalan wajah, pengenalan iris mata, pengenalan bunga, dan pengenalan rambu lalu lintas. Manipulasi data adversarial meniru representasi internal gambar target setelah pembelajaran transfer.

Papernot dkk. melatih model pembelajaran pengganti untuk membuat contoh adversarial yang dapat ditransfer antara beberapa model pembelajaran mendalam dan pembelajaran mesin. Pengklasifikasi yang ditargetkan ditetapkan sebagai oracle yang menggunakan pengambilan sampel reservoir untuk memberi label pada kumpulan data pelatihan dan pada gilirannya meningkatkan efisiensi untuk melatih pengganti pada kumpulan data yang ditambah. Iterasi augmentasi bergantian antara augmentasi kumpulan data yang digunakan untuk melatih model pengganti dan pelabelan yang diberikan oleh oracle untuk menyempurnakan augmentasi. Pengambilan sampel reservoir memengaruhi kualitas pengganti dengan membatasi jumlah kueri pelabelan yang dibuat secara acak dari pengganti ke oracle.

Metode pengambilan sampel seperti ini cocok untuk lingkungan dunia nyata di mana musuh dibatasi oleh kuota untuk terdeteksi oleh pembela HAM. Transferabilitas sampel adversarial dipelajari antara jaringan saraf dalam, regresi logistik, mesin vektor dukungan, pohon keputusan, tetangga terdekat, dan ansambel pengklasifikasi. Bahkan pengklasifikasi pembelajaran mesin komersial yang dihosting oleh Amazon dan Google dipertimbangkan dalam evaluasi eksperimental. Pengganti dirancang untuk serangan kotak hitam di mana musuh menargetkan pengklasifikasi jarak jauh tanpa pengetahuan tentang arsitektur model, parameter, dan kumpulan data pelatihan. Eksperimen menunjukkan bahwa transfer pengetahuan terjadi antara banyak model pembelajaran mesin ke jaringan saraf dalam yang meniru batasan keputusan pengklasifikasi asli.

Untuk mempelajari pembelajaran transfer dengan label target, Liu et al. membedakan antara contoh adversarial yang tidak ditargetkan dan contoh adversarial yang ditargetkan. Contoh adversarial dihasilkan untuk ditransfer ke label target tertentu yang salah diklasifikasikan oleh model pembelajaran mendalam. dagu dkk. mengusulkan metode pembelajaran transfer baru untuk menyempurnakan ketahanan jaringan saraf yang ditransfer yang diperoleh dari mengatur model yang telah dilatih sebelumnya dalam pembelajaran mendalam. Musuh memiliki akses ke bobot dan arsitektur model yang telah dilatih sebelumnya, tetapi tidak memiliki akses ke model dan kueri yang ditransfer khusus tugas.

Baluja dkk. melatih jaringan saraf yang disebut jaringan transformasi adversarial (ATN) untuk membuat serangan adversarial yang ditargetkan. Alih-alih menyelesaikan masalah pengoptimalan per sampel untuk membuat data manipulatif, ATN membuat contoh adversarial yang dimodifikasi secara minimal untuk setiap gambar pelatihan masukan. ATN mengakomodasi berbagai model ancaman seperti melatih target kotak hitam dan kotak putih pada skenario serangan bertarget dan tidak bertarget berdasarkan urutan peringkat dalam keluaran jaringan saraf target. ATN lebih lanjut dapat dilatih untuk menghasilkan gangguan adversarial dari varian jaringan sisa atau pengkodean otomatis adversarial dari masukan yang direkonstruksi dengan sinyal derau adversarial.

Wu dkk. mengidentifikasi contoh manipulasi yang dapat ditransfer karena koneksi lewati dalam pembelajaran mendalam yang diawasi. Gradien dari koneksi lewati diusulkan untuk membuat contoh adversarial. Mereka mentransfer ke jaringan saraf dalam yang canggih termasuk ResNets, DenseNets, Inceptions, Inception-ResNet, dan Squeeze-and-Exitation Networks. Lebih jauh lagi, contoh-contoh adversarial tersebut dapat dikombinasikan dengan teknik kotak hitam yang ada untuk serangan adversarial guna mendapatkan perbaikan dalam metode transferabilitas yang canggih. Contoh-contoh adversarial seperti itu meningkatkan kekhawatiran keamanan dalam penerapan jaringan saraf dalam dalam aplikasi seperti pengenalan wajah, mengemudi otonom, analisis video, dan diagnosis medis.

Contoh manipulasi pembelajaran mesin dalam Adaptasi Domain

Su dkk. mengusulkan adaptasi domain manipulasi dengan pembelajaran aktif. Pengambilan sampel penting yang dikombinasikan dengan pelatihan adversarial digunakan untuk memperhitungkan pergeseran distribusi antar domain. Ini bertindak sebagai skema pemilihan sampel untuk pembelajaran aktif terutama ketika domain target tidak memiliki banyak contoh berlabel seperti domain sumber. Dalam pengambilan sampel kepentingan, keragaman sampel dihasilkan dengan bantuan kerugian adversarial. Adaptasi domain semi-supervisi seperti itu meningkatkan kinerja klasifikasi dan mengurangi biaya pelabelan dengan pembelajaran manipulasi domain pada tugas klasifikasi dan deteksi objek.

Zhang dkk. memperluas adaptasi domain tanpa pengawasan dalam segmentasi semantik dengan pembelajaran manipulatif. Sampel beranotasi tingkat piksel di domain sumber digunakan untuk mengelompokkan sampel tak berlabel di domain target. Dalam pembelajaran manipulasi, diskriminator dibangun untuk membedakan antara domain sumber dan domain target. Model segmentasi kemudian menargetkan untuk menipu diskriminator domain dengan pembelajaran mendalam. Tugas segmentasi semantik adalah memberikan label kelas pada semua piksel dalam suatu gambar. Segmentasi semantik berfungsi sebagai tulang punggung sistem visi komputer seperti kendaraan otonom yang beroperasi di lingkungan perkotaan. Vu dkk. mengatasi tugas adaptasi domain tanpa pengawasan dalam segmentasi semantik dengan kerugian berdasarkan minimalisasi entropi dari prediksi piksel. Pelatihan pembelajaran manipulasi menganalisis jaringan sisa untuk segmentasi semantik guna membangun peta fitur pada domain sumber yang serupa dengan domain target. Jaringan saraf untuk segmentasi semantik dipelajari pada gambar yang dihasilkan dengan konten domain sumber dan gaya domain target yang peta segmentasi sumbernya berfungsi sebagai kebenaran dasarnya. Beberapa paradigma pembelajaran semi-supervisi untuk adaptasi domain dapat memperoleh manfaat dari desain kerugian yang merugikan tersebut.

Yang dkk. mempelajari adaptasi domain dalam segmentasi semantik. Pembelajaran adversarial digunakan untuk mencocokkan distribusi marjinal representasi fitur di seluruh domain. Tujuan serangan diusulkan pada peta fitur perantara yang mempelajari representasi diskriminatif tugas domain-invarian. Mereka diawasi oleh segmentasi semantik di domain sumber. Dengan meningkatkan ketahanan pembelajaran yang diawasi, contoh adversarial yang dapat ditransfer mengisi kesenjangan antar domain dari adaptasi dalam batasan

keputusan klasifikasi. Pembelajaran adversarial seperti itu juga dapat dipahami sebagai bentuk pembelajaran aktif atau penambangan contoh keras yang modelnya meminimalkan kesalahan terburuk ketika fitur diganggu oleh musuh. Fitur adversarial dihasilkan dengan mengumpulkan peta gradien dari tujuan serangan dalam pengklasifikasi segmentasi semantik. Peta prediksi fitur manipulasi yang dimaksudkan untuk membingungkan pengklasifikasi segmentasi dioptimalkan lebih lanjut sesuai dengan teknik minimalisasi entropi yang memberikan pengawasan ekstra dalam tujuan pelatihan.

Kim dkk. merumuskan kembali fungsi pemetaan dalam adaptasi domain untuk menerjemahkan gambar dari satu domain visual ke domain visual lainnya sebagai masalah pembuatan gambar bersyarat untuk jaringan adversarial generatif (GAN). DiscoGAN yang diusulkan tidak memerlukan label eksplisit pada gambar yang dibuat. Hilangnya rekonstruksi gambar diusulkan untuk mendorong pemetaan antara domain gambar multi-modal. Arsitektur GAN baru diberikan untuk mendefinisikan hubungan lintas domain dengan mengubah atribut tertentu seperti warna rambut, jenis kelamin, dan orientasi. Jadi GAN dapat menghasilkan gambar objek dalam adaptasi domain berdasarkan karakteristik, gaya, dan sudut pandang gambar tertentu. Deskripsi teks gambar yang dikodekan dapat digunakan sebagai informasi bersyarat untuk menghasilkan gambar. Modul pengenalan wajah terlatih juga dapat digunakan sebagai masukan bersyarat untuk GAN.

Sankaranarayanan dkk. memberikan pembelajaran embedding untuk adaptasi domain tanpa pengawasan. Hal ini kuat terhadap pergeseran distribusi antara domain sumber dan target. Data tanpa pengawasan dari distribusi target diambil sampelnya untuk memandu prosedur pembelajaran yang diawasi dalam data yang diambil sampelnya dari distribusi sumber. Pendekatan pembuatan gambar adversarial mempelajari penyematan fitur dengan kehilangan klasifikasi dan prosedur pembuatan gambar. Pendekatan yang diusulkan memberikan hasil yang lebih baik daripada penyematan fitur berdasarkan denoising autoencoder dan pengklasifikasi domain. Dengan demikian, kerugian manipulasi dapat melakukan adaptasi domain.

Mancini dkk. menemukan domain laten untuk adaptasi domain. Mereka tertanam dalam arsitektur CNN untuk mempelajari pengklasifikasi target yang kuat. Informasi keanggotaan domain menyelaraskan distribusi representasi fitur CNN dengan distribusi referensi. Pengklasifikasi tersebut mencakup beberapa distribusi domain tanpa memerlukan data pelatihan berlabel. Domain laten tersebut mewakili keragaman gambar domain sumber dan mempelajari informasi tentang kategori semantiknya. Tzeng dkk. menggabungkan pemodelan diskriminatif dengan kerugian GAN untuk menangani pergeseran distribusi yang lebih besar yang tidak ditangani oleh GAN saja. Metode adaptasi adversarial tersebut berupaya meminimalkan perkiraan jarak perbedaan domain melalui tujuan adversarial untuk diskriminator domain. Pekerjaan mengenai kerugian yang merugikan untuk adaptasi domain ini mengasumsikan pilihan desain sebelumnya yang dibuat dalam pembelajaran mendalam untuk adaptasi domain. Ekstraksi fitur khusus domain diperbolehkan untuk dipelajari dengan tidak membagikan bobot jaringan saraf antara domain sumber dan target. Pengklasifikasi

domain target dilatih secara berlawanan hingga cocok dengan prediksi pengklasifikasi domain sumber. Aplikasi didemonstrasikan pada tugas adaptasi lintas modalitas.

Shen dkk. mempelajari representasi domain-invarian dengan pembelajaran representasi berpemandu jarak jauh (WDGRL) Wasserstein. Dengan memberikan tujuan adversarial pada pengklasifikasi domain, permainan min-maks dirancang untuk adaptasi domain agar representasi fitur sumber dan target tidak dapat dibedakan. Pengklasifikasi domain membedakan antara representasi sumber dan target di mana jarak Wasserstein bertindak sebagai ukuran perbedaan domain untuk kerugian adversarial. WDGRL dioptimalkan dengan strategi pelatihan adversarial berulang untuk meminimalkan perkiraan jarak Wasserstein antara representasi fitur sumber dan target. Oleh karena itu, pembelajaran mendalam bertindak sebagai kerangka kerja yang kuat untuk mempelajari representasi fitur untuk adaptasi domain. Secara spesifik, jarak Wasserstein mampu menghubungkan kesalahan sumber dan target.

Wang dkk. mengusulkan fungsi kerugian tujuan adversarial untuk menjembatani domain sumber dan target dengan mempelajari representasi mendalam domain-invarian pada wilayah yang dapat ditransfer dalam gambar. Pembelajaran transfer seperti itu mampu menghasilkan model diskriminatif yang mengurangi pergeseran kumpulan data antara distribusi pelatihan dan pengujian. Representasi domain-invarian yang dihasilkan dapat ditanamkan ke dalam arsitektur jaringan neural dalam untuk meminimalkan perbedaan antara distribusi fitur sumber dan target dengan pembelajaran adversarial. Mekanisme perhatian dalam pembelajaran mendalam disorot untuk mengekstraksi fitur-fitur terperinci dengan mempertimbangkan berbagai wilayah gambar yang diperoleh dari domain berbeda. Mekanisme perhatian lokal yang dapat dialihkan diusulkan untuk menghasilkan beberapa diskriminator domain tingkat wilayah, dan mekanisme perhatian global yang saling melengkapi diusulkan untuk menghasilkan diskriminator domain tingkat gambar tunggal untuk menyorot gambar yang dapat ditransfer.

Lebih lanjut, pelatihan model manipulasi dapat diperluas ke beberapa diskriminator untuk meningkatkan pencocokan distribusi dimana desain diskriminator dapat berkisar dari musuh yang tangguh hingga guru yang pemaaf. Di sini, permainan min-maks diusulkan antara diskriminator dan generator sehingga diskriminator domain membedakan antara data sumber dan target, sedangkan generator adalah ekstraktor fitur yang dilatih secara bermusuhan untuk menipu diskriminator domain. Desain dalam adaptasi domain mendalam ini merupakan perluasan dari ide klasik dalam statistik untuk menentukan jarak statistik dalam ruang metrik probabilistik. Jarak tersebut kemudian diminimalkan dengan mempelajari representasi data sumber dan target yang mampu menjembatani kesenjangan distribusi antar domain yang berbeda. Model perhatian yang dapat dialihkan berdasarkan pada kerugian yang merugikan dalam mentransfer perhatian antara objek di domain sumber dan target memiliki aplikasi dalam keterangan gambar, segmentasi gambar, dan klasifikasi gambar.

Pengklasifikasi jaringan dalam yang dihasilkan dapat dilatih pada domain data sumber berlabel dan menggeneralisasi dengan baik ke domain data target yang tidak berlabel.

Mereka dapat disempurnakan lebih lanjut menuju adaptasi pengklasifikasi dengan meminimalkan entropi distribusi bersyarat kelas untuk diatur pada domain target. Mekanisme pembangkitan perhatian lokal menciptakan nilai entropi penuh perhatian untuk setiap kehilangan entropi gambar guna meningkatkan pencocokan gambar serupa di seluruh domain sumber dan target. Hilangnya entropi penuh perhatian dikombinasikan dengan tujuan adaptasi domain dan tujuan klasifikasi adversarial untuk mendapatkan masalah terpadu yang dioptimalkan untuk pelatihan adversarial. Solusi optimal untuk membuat representasi fitur kemudian diperoleh dengan prosedur backpropagation pada error yang dapat dihitung pada differentialiable loss. Representasi fitur tersebut dapat diperluas ke masalah pengoptimalan multimodal yang melibatkan pemilihan fitur yang berlawanan dalam pembelajaran mesin yang tangguh.

Contoh Adversarial dalam Domain Keamanan Siber

Contoh adversarial pertama kali dibuat dalam klasifikasi gambar. Karena kedalaman arsitektur dalam pengklasifikasi pembelajaran mendalam seperti CNN, kemampuan interpretasi jutaan parameter yang dipelajari dalam model tersebut menjadi sangat penting. Hal ini telah diperluas ke mekanisme yang lebih kompleks untuk menyerang pengenalan wajah, pengenalan tindakan video, dan serangan adversarial dunia fisik terhadap rambu-rambu jalan. Wei dkk. menghasilkan contoh adversarial yang dapat ditransfer untuk deteksi objek gambar dan video. Kemampuan transfer contoh adversarial ditingkatkan dengan memanipulasi peta fitur tingkat rendah dari beberapa lapisan pendeteksi objek. Mekanisme pembobotan perhatian diintegrasikan ke dalam hilangnya fitur untuk memanipulasi subkawasan fitur. Kerugian kelas tingkat tinggi digunakan untuk melatih generator. Pembuatan contoh adversarial dirumuskan sebagai masalah penerjemahan gambar-ke-gambar. Contoh adversarial tersebut dapat dibuat untuk dua jenis model umum untuk deteksi objek yang dikategorikan sebagai model berbasis proposal dan model berbasis regresi.

Contoh adversarial dapat dibuat dengan mengubah nilai tingkat piksel dalam klasifikasi gambar. Mereka juga telah dibuat dengan menerapkan patch perubahan pada gambar yang digunakan dalam deteksi objek pada tanda berhenti, misalnya. Ini dkk. membuat tambalan manipulatif pada detektor orang. Di sini, kelas target berisi banyak variasi intrakelas, tidak seperti kumpulan data tanda berhenti. Serangan manipulatif tersebut dapat digunakan sebagai perangkat penyelubungan untuk menghindari sistem pengawasan di mana penyusup dapat menyelinap tanpa terdeteksi dengan memegang pelat karton kecil di depan tubuh mereka yang diarahkan ke kamera pengintai. Mereka dapat menambah gambar yang diberi anotasi manusia untuk menentukan performa model untuk deteksi orang. Rangkaian pengujian tersebut memperhitungkan contoh-contoh adversarial yang dirancang untuk mengarahkan model ke arah yang salah dan menargetkan lebih lanjut untuk mengelabui model.

Kerentanan dalam model pendeteksian orang pada kamera pengintai keamanan dapat disorot sebagai risiko serangan terhadap sistem pendeteksian. Kotak pembatas untuk patch adversarial diprediksi berdasarkan skor objek dan komponen skor kelas dalam kekalahan adversarial. Patch adversarial kemudian diterapkan pada gambar setelah berbagai

transformasi untuk lebih mengelabui detektor. Hal ini memungkinkan terjadinya serangan yang ditargetkan di mana data tersedia untuk adegan tertentu di lingkungan rekaman. Beberapa faktor yang mempengaruhi timbulnya patch adversarial adalah perubahan pencahayaan, perbedaan sudut pandang, rotasi patch, dan ukuran patch. Mereka dapat berubah sesuai dengan ukuran orang, kamera dapat menambahkan noise atau mengaburkan patch. Mereka mengoptimalkan gambar untuk meminimalkan berbagai kemungkinan terkait kemunculan seseorang pada keluaran detektor. Dalam eksperimen, efek patch yang dihasilkan dibandingkan dengan patch acak untuk menentukan patch paling efektif yang meminimalkan kehilangan objek. Mengoptimalkan kerugian adversarial untuk arsitektur detektor yang berbeda memastikan transferabilitas patch adversarial.

Elsayed dkk. membuat contoh adversarial yang ditransfer dari model visi komputer ke pengamat manusia yang memiliki waktu terbatas. Pengaruh contoh adversarial dalam pembelajaran mesin diselidiki berbeda dengan bias kognitif dan ilusi optik dalam persepsi visual manusia yang dipelajari oleh ilmu saraf. Jadi, dimungkinkan untuk membuat contoh-contoh adversarial dengan fitur-fitur yang bermakna bagi manusia. Mereka dapat dirancang untuk menyebabkan kesalahan tidak hanya dalam pengenalan objek visual tetapi juga dalam persepsi manusia. Elsayed dkk. merancang eksperimen psikofisika untuk membandingkan pola kesalahan yang dibuat oleh manusia dengan validasi kesalahan klasifikasi dalam pengklasifikasi jaringan saraf.

Coklat dkk. membuat patch gambar adversarial yang ditargetkan yang dapat menyerang pemandangan apa pun di dunia fisik untuk menyebabkan pengklasifikasi gambar mengeluarkan kelas target apa pun dalam berbagai transformasi matematika. Pengetahuan sebelumnya tentang kondisi pencahayaan, sudut kamera, dan tipe pengklasifikasi yang menjadi target tidak diperlukan untuk menciptakan serangan dunia fisik seperti itu. Dalam tugas klasifikasi citra, pengklasifikasi harus mendeteksi fitur yang paling menonjol dalam suatu citra untuk menentukan label kelasnya. Jalur adversarial memanfaatkan fitur ini untuk menghasilkan fitur adversarial yang jauh lebih menonjol dibandingkan objek di dunia fisik. Gangguan lokal yang begitu besar dan tidak terlihat juga dapat menyesatkan pengklasifikasi pembelajaran mesin yang beroperasi tanpa validasi manusia. Jadi contoh-contoh adversarial dapat dibuat untuk dunia fisik dengan memodelkan contoh-contoh adversarial dari transformasi fisik di mana robot mengamati dunia melalui kamera, sensor, dan telepon untuk menangani representasi data gambar, suara, dan video.

Athalye dkk. mensintesis objek adversarial 3D yang berlawanan dengan distribusi transformasi yang dipilih seperti pergeseran sudut pandang, kebisingan kamera, dan transformasi affine. Algoritme ekspektasi atas transformasi dirancang dalam skenario serangan kotak putih di mana musuh memiliki akses ke pengklasifikasi, gradiennya, kelas yang memungkinkan, dan ruang masukan yang valid. Dalam prosedur optimasi untuk membuat contoh adversarial, gangguan adversarial dimodelkan sehubungan dengan ekspektasi yang ditentukan pada distribusi fungsi transformasi yang dipilih. Alih-alih memilih kemungkinan log dari satu contoh sebagai tujuan pengoptimalan, jarak efektif antara masukan yang berlawanan dan masukan asli diminimalkan. Ini adalah jarak yang diharapkan atau dirasakan

seperti yang dilihat oleh pengklasifikasi. Solusi optimal yang menghasilkan data adversarial diperoleh dengan algoritma penurunan gradien stokastik dari nilai yang diharapkan dimana gradien dihitung melalui diferensiasi melalui setiap transformasi pengambilan sampel. Skenario serangan manipulatif seperti itu memperlakukan dunia siber sebagai sebuah domain yang transformasinya ditransfer ke dunia fisik dan bertindak sebagai sebuah kodomain. Distribusi transformasi bertindak sebagai anggaran perturbasi untuk menghasilkan contoh-contoh manipulasi yang berhasil.

Sistem pembelajaran mesin rentan terhadap serangan adversarial, terutama di lingkungan adversarial non-stasioner dalam domain keamanan siber. Di luar domain pengenalan gambar seperti deepfake dalam sistem deteksi, aplikasi pembelajaran adversarial dalam domain keamanan siber mencakup identifikasi malware, deteksi spam, penilaian risiko, injeksi SQL, pengembangan ransomware, sistem pengenalan biometrik, deteksi rambu lalu lintas, mengemudi secara otonom, deteksi anomali, klasifikasi entitas , pembelajaran kamus, sistem cyber-fisik, dan sistem kontrol industri. Dalam domain keamanan siber, memodifikasi panggilan API atau byte konten yang dapat dieksekusi dapat menyebabkan executable yang dimodifikasi tersebut menjalankan fungsi yang berbeda. Jadi musuh di domain keamanan siber harus menerapkan metode untuk memodifikasi fitur yang dapat dieksekusi agar tidak merusak fungsinya karena sampel data yang terganggu dalam vektor fitur dalam karakter URL, email spam, paket jaringan, pendeteksi phishing, sinyal sensor, proses fisik, dll.

Beberapa salah satu serangan yang ditargetkan pada jaringan saraf yang dibangun untuk domain keamanan siber adalah serangan APT khusus, serangan Trojan, serangan pintu belakang, dan serangan penolakan layanan terdistribusi (DDoS). Dalam sistem cyber-fisik, aplikasi pembelajaran adversarial berperan dalam optimalisasi infrastruktur penting seperti jaringan tenaga listrik, jaringan transportasi, jaringan pasokan air, dan pembangkit listrik tenaga nuklir. Dalam sistem pengenalan biometrik, pembelajaran manipulasi memiliki aplikasi dalam verifikasi tanda tangan tulisan tangan, klasifikasi sidik jari, pengenalan wajah, analisis sentimen, pengenalan pembicara, forensik jaringan, dan pembuatan kode iris mata.

BAB 3
PERMUKAAN SERANGAN MUSUH

Dalam bab ini, kita mengeksplorasi permukaan serangan yang bermusuhan. Kami memeriksa bagaimana mereka dapat mengeksploitasi kerentanan dalam pembelajaran mesin dan bagaimana membuat algoritma pembelajaran kuat terhadap serangan terhadap keamanan dan privasi sistem pembelajaran. Untuk mengeksplorasi kerentanan, kami dapat mensimulasikan berbagai proses pelatihan model dalam berbagai skenario serangan dalam pengaturan yang diawasi dan tidak diawasi. Setiap strategi serangan diasumsikan dirumuskan oleh musuh cerdas yang mampu melakukan manipulasi fitur, manipulasi label, atau keduanya.

Kebijakan serangan optimal dari musuh ditentukan oleh solusi untuk masalah optimasi yang menghasilkan data musuh. Kami kemudian dapat menerapkan pengetahuan yang kami pelajari untuk meningkatkan dan memperkuat prosedur pembelajaran agar dapat bertahan lebih baik dari serangan. Analisis sensitivitas yang dirangkum dalam bab ini dapat digunakan untuk mengembangkan algoritma komputasi untuk tujuan optimasi dan inferensi statistik dalam kapasitas algoritma pembelajaran adversarial untuk pengacakan, diskriminasi, keandalan, dan kemampuan belajar. Hal ini menciptakan jalur penelitian mengenai ketahanan, keadilan, penjelasan, dan transparansi model pembelajaran mesin.

3.1 KEAMANAN DAN PRIVASI DALAM MANIPULASI PEMBELAJARAN MESIN

Serangan Penghindaran Biggio dkk. membahas keamanan adversarial pada waktu pengujian sistem pengklasifikasi yang diterapkan. Evaluasi keamanan kemudian dilakukan pada tingkat risiko yang berbeda dari kinerja pengklasifikasi non-linier dalam deteksi malware. Pengklasifikasi aman diusulkan dengan menggunakan pendekatan penurunan gradien pada fungsi diskriminan yang dapat dibedakan. Sasaran Adversary didefinisikan dalam bentuk meminimalkan fungsi kerugian pengklasifikasi dengan sampel adversarial positif yang melintasi batas keputusan. Model ini juga dapat menggabungkan pengetahuan adversarial khusus aplikasi dalam definisi skenario serangan adversarial. Pengetahuan adversarial tersebut mencakup pengetahuan sebelumnya tentang data pelatihan, representasi fitur, jenis algoritma pembelajaran dan fungsi keputusannya, bobot klasifikasi, dan umpan balik dari pengklasifikasi.

Serangan Peracunan Dalam pengaturan yang sensitif terhadap keamanan, algoritme pembelajaran mesin tidak dapat berasumsi bahwa data pelatihan berasal dari distribusi yang alami dan berperilaku baik. Dengan memasukkan contoh manipulasi ke dalam data pelatihan sehingga kesalahan pengujian meningkat, Biggio dkk. menyelidiki serangan keracunan terhadap mesin vektor dukungan (SVM) dengan kernel linier, kernel polinomial, dan kernel RBF. Prosedur pendakian gradien digunakan untuk menghitung contoh manipulasi sebagai maksimum lokal dari permukaan kesalahan non-cembung SVM. Dalam iterasi pendakian gradien, setelah setiap pembaruan pada contoh serangan, batas keputusan optimal dihitung

dari solusi ke SVM tambahan. Prosedur pencarian memerlukan banyak langkah gradien kecil. Ini dihentikan ketika contoh serangan menyimpang terlalu banyak dari data pelatihan. Perubahan pada batasan keputusan SVM karena masukan berbahaya terbukti penting dalam domain aplikasi seperti spam, worm, intrusi, dan deteksi penipuan.

Xiao dkk. mengusulkan gangguan label adversarial untuk memaksimalkan kesalahan klasifikasi kasus terburuk SVM dengan membalik label pada data pelatihan. Strategi serangan untuk menciptakan gangguan label yang merugikan dimotivasi oleh kerangka minimalisasi risiko struktural. Dalam kerangka ini, pembelajaran SVM meminimalkan jumlah risiko pengatur dan risiko empiris dalam data. Di sini, pengatur memberikan penalti pada kompleksitas hipotesis yang berlebihan untuk menghindari overfitting dalam masalah pemrograman kuadrat optimasi cembung. Musuh kemudian mengoptimalkan risiko empiris pada data berbahaya sehingga SVM disesatkan sehingga menggeser batasan keputusan dari distribusi data asli. Optimalisasi risiko empiris selanjutnya didekomposisi menjadi dua sub-masalah berulang yang diselesaikan dengan pemrograman kuadratik dan pemrograman linier. Label sampel lebih lanjut di kelas yang berbeda dibalik dengan cara yang berkorelasi untuk memaksa batas keputusan pembentuk hyperplane berputar sebanyak mungkin. Musuh diasumsikan memiliki pengetahuan penuh tentang kumpulan fitur dalam data pelatihan dengan biaya yang sama yang ditetapkan untuk setiap pembalikan label.

Serangan Inferensi Shokri dkk. menyelidiki masalah pelanggaran privasi model klasifikasi komersial yang membocorkan informasi tentang data pelatihan mereka pada aplikasi Internet. Kueri musuh menargetkan model sebagai kotak hitam untuk mengambil keluaran model pada masukan tertentu. Masukan tersebut dihasilkan dengan melatih model bayangan yang meniru perilaku model target. Berbeda dengan model target kotak hitam, model bayangan mengetahui label kebenaran dasar untuk catatan yang disimpulkan. Model kotak hitam adalah model jaringan saraf di Amazon ML dan Google Prediction API yang dilatih pada kumpulan data gambar. Detail model kotak hitam disembunyikan dari pemilik datanya. Dataset diperoleh dari pembelian retail, penelusuran lokasi, dan rawat inap di rumah sakit. Di sini, pelanggaran privasi dikatakan terjadi jika musuh dapat menggunakan keluaran model untuk menyimpulkan nilai atribut sensitif dalam masukan model.

Inferensi atribut didefinisikan oleh Shokri et al. dalam hal inferensi keanggotaan kelas dari keberadaan catatan data tertentu dalam kumpulan data pelatihan model. Keberhasilan inferensi keanggotaan kelas yang diusulkan diukur dalam hal presisi serangan dan ingatan serangan model target. Model bayangan dilatih pada kumpulan data sintetis dengan algoritma pendakian bukit yang menghasilkan catatan kandidat yang diklasifikasikan dengan keyakinan tinggi oleh model target. Dalam setiap iterasi pendakian bukit, sebuah kandidat rekaman diusulkan dengan mengubah fitur-fitur yang dipilih secara acak dari rekaman terbaru yang diterima. Catatan kandidat diterima dalam algoritma pendakian bukit hanya jika catatan tersebut meningkatkan kemungkinan diklasifikasikan dengan benar berdasarkan model target. Beberapa strategi pertahanan diusulkan terhadap pertanyaan keanggotaan kelas. Strategi ini termasuk membatasi vektor prediksi ke kelas atas, membulatkan probabilitas klasifikasi, meningkatkan entropi vektor prediksi sehingga keluaran menjadi hampir seragam

dan tidak bergantung pada masukan, dan mengatur fungsi kerugian klasifikasi untuk menghukum parameter besar selama pelatihan.

Serangan Pengklasifikasi Linier

Dalvi dkk. menganalisis kinerja pengklasifikasi dengan melihat klasifikasi sebagai permainan antara pengklasifikasi yang beradaptasi dengan musuh yang berusaha membuat pengklasifikasi menghasilkan negatif palsu. Di sini, musuh yang sensitif terhadap biaya dikontraskan dengan pengklasifikasi yang sensitif terhadap biaya di mana proses menghasilkan data dalam klasifikasi adversarial tidak hanya diperbolehkan untuk berubah seiring waktu namun juga memungkinkan perubahan ini menjadi fungsi dari parameter pengklasifikasi. Klasifikasi adversarial didefinisikan sebagai permainan antara dua pemain musuh dan pengklasifikasi di mana pengklasifikasi memaksimalkan fungsi imbalannya yang dicirikan oleh ekspektasi imbalan pengklasifikasi atas parameter biaya musuh.

Dalvi dkk. mengusulkan bahwa tujuan musuh adalah menemukan strategi perubahan fitur klasifikasi yang memaksimalkan hasil yang diharapkan musuh. Contoh adversarial dihasilkan oleh algoritma pemilihan fitur standar dengan fungsi pembayaran pengklasifikasi Naive Bayes sebagai fungsi evaluasi. Teori strategi keseimbangan Nash yang dapat diatur secara komputasi dalam klasifikasi adversarial dibiarkan sebagai pertanyaan terbuka yang menganalisis permainan non-zero sum dua pemain. Lowd dkk. memperkenalkan algoritma pembelajaran manipulasi untuk pengklasifikasi linier yang diserang. Tujuan dari pembelajaran adversarial adalah untuk mempelajari dan menyerang bagian dari batasan keputusan pengklasifikasi dengan mempelajari bobot fitur tanpa (i) membuat representasi fitur spesifik domain dan (ii) mengasumsikan proses stokastik untuk distribusi data pelatihan.

Lowd dkk. berasumsi bahwa musuh dapat mengirimkan pertanyaan keanggotaan ke pengklasifikasi untuk membedakan antara contoh berbahaya dan contoh tidak berbahaya. Kompleksitas komputasi dari kemungkinan kueri keanggotaan dibatasi oleh sejumlah polinomial penelusuran garis di sepanjang setiap dimensi fitur. Algoritme pembelajaran adversarial kemudian meminimalkan fungsi biaya adversarial linier pada ruang instance non-berbahaya untuk dipelajari oleh musuh. Fungsi biaya adversarial optimal menghasilkan contoh tidak berbahaya yang paling mirip dengan contoh dasar adversarial yang dapat diakses oleh musuh.

Lowd dkk. juga mendemonstrasikan eksperimen pelatihan adversarial pada pengklasifikasi linier seperti model Naive Bayes, mendukung mesin vektor dengan kernel linier, dan model entropi maksimum yang mempelajari fitur Boolean untuk pemfilteran spam. Kerangka pembelajaran yang diusulkan (disebut ACRE) berguna untuk mempelajari penyerang atau musuh dan pembela atau pengklasifikasi. Hal ini dapat digunakan untuk menentukan apakah musuh dapat belajar secara efisien dalam mengalahkan pengklasifikasi dengan meminimalkan fungsi biaya adversarial.

3.2 FITUR SERANGAN PEMBOBOTAN

Secara tradisional, algoritme pembelajaran mesin berasumsi bahwa pelatihan algoritme dapat dilakukan pada data yang terkontrol dan berkualitas tinggi. Di dunia nyata,

pembelajaran mesin dilakukan pada data yang berisik dan tidak pasti. Di sini, pengklasifikasi yang kuat dapat mengantisipasi fitur-fitur berisik selama pengujian hanya ketika pengklasifikasi tersebut dilatih dengan asumsi fitur-fitur berisik ada selama pelatihan. Selain itu, pengklasifikasi yang kuat harus berupa pengklasifikasi padat yang melatih sebanyak mungkin fitur informatif atau penting. Pertimbangan tersebut adalah fokus dari teknik pembobotan fitur di lingkungan yang bermusuhan. Di sini, data adversarial yang diciptakan oleh musuh yang cerdas berbeda dari gangguan acak yang ditemukan di alam.

Karena pengklasifikasi tradisional tidak dapat terus-menerus menyesuaikan diri terhadap perubahan lingkungan yang berlawanan, Kolcz dkk. mencoba merancang pengklasifikasi yang terdegradasi dengan baik karena distribusi data pengujian berbeda dari distribusi data pelatihan asli. Hal ini dilakukan dengan proses pemilihan fitur yang membobot ulang fitur-fitur yang kurang penting untuk klasifikasi. Pembobotan fitur meningkatkan performa model dengan menjadikannya kuat terhadap penyimpangan konsep dalam data dengan mengorbankan biaya komputasi tambahan dalam model. Intuisi di balik pendekatan ini adalah bahwa distribusi bobot pada fitur-fitur untuk algoritma pembelajaran mencerminkan pentingnya fitur-fitur untuk pembelajaran tanpa pengawasan dan pengawasan.

Kolcz dkk. membayangkan pendekatan dua tahap untuk pelatihan pengklasifikasi yang kuat di mana pengklasifikasi digunakan untuk menetapkan bobot pada fitur di tahap pertama yang kemudian diubah melalui pembobotan fitur untuk menginduksi model akhir di tahap kedua. Model akhir memenuhi fungsi tujuan untuk dioptimalkan. Karena skema pembobotan ulang terbaik untuk pembelajaran yang diawasi dan tidak diawasi tidak tersedia dalam literatur, Kolcz dkk. bereksperimen dengan beberapa pilihan untuk pembobotan fitur. Fungsi tujuan untuk pembobotan fitur dianalisis secara formal oleh Kolcz et al. sebagai kasus khusus minimalisasi risiko yang diatur dengan pengatur bentuk kuadrat dan fungsi kerugian cembung.

Metode pembobotan fitur dalam eksperimen Kolcz et al. mencakup pengantongan fitur, regresi logistik terpartisi, pembelajaran berbobot kepercayaan, injeksi gangguan fitur, dan koreksi bias pemilihan sampel. Detailnya dijelaskan di bawah ini:

- Feature bagging melatih model probabilistik sebagai rata-rata aritmatika atau geometrik dari beberapa model dasar. Setiap model dasar dilatih tentang kemungkinan subkumpulan fitur asli yang tumpang tindih. Kinerja model yang dikantongi diharapkan menjadi lebih tangguh dibandingkan performa model dasar mana pun karena bobot fitur yang kurang penting akan digantikan oleh bobot fitur yang lebih penting selama proses pelatihan.
- Regresi logistik terpartisi adalah kasus khusus dari feature bagging dimana subset fitur dan label kelas tidak tumpang tindih.
- Untuk mencegah undertraining, pembelajaran berbobot percaya diri secara agresif memperbarui bobot fitur langka dalam data dengan mempertahankan distribusi normal pada vektor bobot pengklasifikasi linier. Bobot fitur diperbarui sedemikian

rupa sehingga perbedaan Kullback-Leibler antara distribusi data pelatihan dan distribusi data pengujian diminimalkan tanpa mengurangi performa model.

- Injeksi derau fitur mengurangi masalah overfitting model pada data pelatihan dengan memperkenalkan derau fitur buatan selama pelatihan model.
- Koreksi bias pemilihan sampel menetapkan bobot fitur sedemikian rupa sehingga data pelatihan yang diberi bobot ulang menyerupai data pengujian yang tersedia. Bobot yang benar disimpulkan tanpa estimasi kepadatan yang eksplisit. Namun, koreksi bias pemilihan sampel mengasumsikan data pengujian juga tersedia selama pelatihan di domain masukan.

Liu dkk. merancang algoritma pembelajaran yang diawasi yang aman terhadap serangan keracunan yang tidak membuat asumsi independensi pada distribusi fitur. Serangan keracunan diasumsikan pada langkah reduksi dimensi dan regresi prediktif. Fitur berdimensi tinggi diproyeksikan ke subruang berdimensi rendah dengan kepadatan data tinggi. Kemudian model regresi linier karakteristik data terbaik.

Algoritme faktorisasi matriks diusulkan untuk memulihkan subruang berdimensi rendah dengan adanya data pelatihan yang rusak oleh contoh noise dan adversarial. Regresi komponen utama menggunakan pengoptimalan yang dipangkas untuk memperkirakan parameter regresi dalam subruang berdimensi rendah. Dalam skenario serangan adversarial yang diusulkan oleh Liu dkk., model regresi dapat memilih proses pelatihan dan strategi pertahanannya tanpa akses ke data pelatihan sebelum manipulasi adversarial. Musuh memiliki pengetahuan penuh tentang algoritma dan parameter pelatihan. Skenario serangan adversarial disimulasikan sebagai permainan Stackelberg zero-sum di mana fungsi pembayaran musuh meminimalkan anggaran tertentu untuk meracuni data pelatihan, sedangkan fungsi pembayaran regressor adalah akurasi regresi.

Proses pembelajaran secara formal dicirikan dalam bentuk fungsi model yang menghubungkan masukan yang berlawanan dan keluaran yang diprediksi. Fungsi kerugian kuadrat dan fungsi ambang batas fungsi kerugian juga dianalisis dalam regresi. Maksimalisasi alternatif memecahkan masalah pengoptimalan yang diusulkan pada log HTTP. Data adversarial dihasilkan dengan memindahkan sampel data pelatihan sepanjang arah untuk memanipulasi regressor hingga tidak dapat memprediksi dengan benar. Hasilnya dibandingkan dengan model regresi yang kuat seperti regresi linier OLS dan prediksi regresi ridge dengan adanya noise.

3.3 KESALAHAN MESIN VEKTOR PENDUKUNG (SVM)

Menurut mekanisme keamanan adversarial yang telah dibahas sebelumnya, serangan peracunan adalah serangan kausatif di mana titik serangan yang dibuat khusus dimasukkan ke dalam data pelatihan. Dalam serangan keracunan, musuh tidak dapat mengakses database pelatihan namun dapat memberikan data pelatihan baru. Serangan keracunan membahayakan keamanan sistem pembelajaran berskala besar yang menyimpulkan pola tersembunyi dalam kumpulan data besar yang rumit untuk mendukung pengambilan

keputusan dengan statistik perilaku. Serangan keracunan sebelumnya telah dipelajari dengan metode deteksi anomali.

Dalam model yang mungkin kira-kira benar (PAC), minimalisasi risiko struktural pembelajaran SVM dipelajari dalam konteks masalah pemrograman kuadratik cembung. Dampak kebisingan label stokastik dan adversarial pada kesalahan klasifikasi mesin vektor dukungan (SVM) telah dianalisis secara teoritis dalam model pembelajaran PAC. Serangan keracunan di SVM telah diatasi dengan mempertimbangkan sanitasi data sebagai bentuk deteksi outlier, sistem pengklasifikasi ganda, pembelajaran tambahan, dan statistik yang kuat. Serangan penghindaran dalam SVM telah diatasi dengan secara eksplisit menanamkan pengetahuan tentang manipulasi data yang merugikan ke dalam algoritma pembelajaran menggunakan (i) model teoritis permainan untuk klasifikasi, (ii) model probabilistik dari penyimpangan distribusi data yang diserang, dan (iii) sistem pengklasifikasi ganda.

Biggio dkk. menunjukkan bahwa musuh yang cerdas dapat memprediksi perubahan dalam fungsi keputusan SVM karena masukan musuh. Serangan keracunan terhadap SVM memasukkan contoh manipulasi ke dalam data pelatihan untuk meningkatkan kesalahan pengujian SVM. Serangan yang diusulkan oleh Biggio dkk. memiliki teknik pembelajaran tambahan dengan strategi pendakian gradien. Gradien dihitung berdasarkan properti solusi optimal SVM. Karena serangan bergantung pada gradien (perkalian titik antar titik di ruang masukan), serangan juga dikernelisasi dengan menggunakan kernel linier dan non-linier di ruang masukan.

Untuk meningkatkan kesalahan pengujian, prosedur pendakian gradien menyatu ke maksimum lokal dari permukaan kesalahan validasi non-cembung. Strategi pendakian gradien yang diusulkan mengasumsikan bahwa musuh mengetahui data pelatihan yang digunakan oleh algoritma pembelajaran. Dalam skenario serangan di dunia nyata, kumpulan data pelatihan pengganti dapat digunakan sebagai ganti kumpulan data pelatihan asli. Konvergensi strategi pendakian gradien yang diusulkan bergantung pada kelancaran parameter SVM dan manifold geometri titik data yang ditemukan dalam solusi masalah pemrograman kuadrat. Strategi serangan yang diusulkan juga dapat diperluas ke koalisi serangan di mana pemilihan subset titik data terbaik untuk serangan merupakan masalah pemilihan subset.

Huang dkk. mengusulkan algoritma pembelajaran adversarial untuk serangan terhadap SVM yang memaksimalkan kesalahan klasifikasi dengan membalik label pada data pelatihan. Serangan kontaminasi yang diusulkan adalah serangan keracunan karena menargetkan kesalahan pengujian SVM (juga disebut risiko empiris dalam model PAC) dengan mengkontaminasi label data pelatihan. Manipulasi data adversarial yang diusulkan disebut injeksi kebisingan label. Dua algoritma serangan diusulkan untuk memperhitungkan manipulasi data yang merugikan. Kedua algoritme tersebut berasumsi bahwa musuh memiliki akses ke kumpulan fitur data pelatihan. Setiap label yang dibalik oleh musuh diasumsikan memiliki biaya yang sama dan tidak bergantung pada nilai fitur dalam sampel. Algoritme pertama dengan rakus memaksimalkan kesalahan pengujian SVM melalui pelonggaran nilai label secara terus menerus dalam prosedur pendakian gradien.

Algoritme kedua melakukan penelusuran luas pertama untuk dengan rakus membuat kumpulan kandidat label yang dibalik yang berkorelasi dengan kesalahan pengujian SVM. Kedua algoritma tersebut dapat dipahami sebagai pencarian label yang mencapai perbedaan maksimum antara risiko empiris untuk pengklasifikasi yang dilatih pada data asli dan data terkontaminasi. Algoritme ini juga dapat digunakan untuk mensimulasikan permainan penjumlahan konstan antara penyerang dan pengklasifikasi yang bertujuan untuk memaksimalkan dan meminimalkan kesalahan pengujian pada kumpulan data pengujian yang tidak ternoda. Formulasi permainan yang berbeda dapat disimulasikan jika pemain menggunakan fungsi tujuan non-antagonis. Perbaikan algoritma dimungkinkan dengan mempelajari SVM tambahan di bawah gangguan label. Masalah injeksi noise label yang menciptakan manipulasi penyerang di SVM juga terkait dengan masalah klasifikasi SVM dalam pembelajaran semi-supervised, pembelajaran aktif, dan prediksi terstruktur.

3.4 ANSAMBEL PENGKLASIFIKASI YANG KUAT

Biggio dkk. mengusulkan bahwa kumpulan pengklasifikasi linier tidak hanya dapat meningkatkan akurasi tetapi juga ketahanan pembelajaran yang diawasi. Hal ini karena lebih dari satu pengklasifikasi harus dihindari atau diracuni untuk membahayakan seluruh kelompok pengklasifikasi. Strategi pelatihan mendistribusikan bobot fitur secara merata antara fitur diskriminatif dan non-diskriminatif dalam data. Meremehkan bobot diskriminatif dalam pengklasifikasi dapat melemahkan keakuratan pengklasifikasi. Tujuan dari ansambel pengklasifikasi yang kuat adalah untuk menemukan trade-off yang tepat antara ketahanan dan akurasi. Di sini, musuh dipaksa untuk mengubah sejumlah besar nilai fitur untuk memanipulasi pengklasifikasi.

Biggio dkk. merancang metode boosting dan random subspace (RSM) untuk mendistribusikan bobot dalam algoritma adversarial. Perilaku adversarial dimodelkan dalam dua skenario—skenario terburuk di mana musuh memiliki pengetahuan lengkap tentang pengklasifikasi dan skenario kasus rata-rata di mana musuh hanya memiliki perkiraan pengetahuan tentang pengklasifikasi. Fungsi diskriminasi ansambel kemudian diperoleh dengan membuat rata-rata pengklasifikasi linier berbeda yang dilatih pada subset berbeda yang dipilih secara acak dari kumpulan fitur asli.

Metode rata-rata oleh Biggio et al. untuk menemukan kinerja ansambel merupakan perpanjangan dari gagasan untuk menggunakan kinerja rata-rata pengklasifikasi linier untuk mencegah overfitting atau underfitting pada data yang tidak seimbang. Dengan mengurangi komponen varians kesalahan klasifikasi atau estimasi, pengambilan sampel secara acak yang digunakan dalam algoritma mengurangi ketidakstabilan dalam fungsi pengambilan keputusan atau estimasi. Fungsi keputusan yang stabil seperti itu tidak seharusnya mengalami perubahan besar pada keluaran karena gangguan kecil pada data masukan karena data adversarial atau gangguan stokastik.

Dalam Biggio dkk., evaluasi eksperimental memiliki dua tujuan. Tujuan pertama adalah untuk memahami kondisi di mana pengambilan sampel secara acak menghasilkan distribusi bobot yang merata dalam kumpulan pengklasifikasi. Tujuan kedua adalah untuk

mengevaluasi apakah bobot yang didistribusikan secara merata meningkatkan ketahanan ansambel pengklasifikasi dibandingkan dengan pengklasifikasi dasar tunggal. Dengan demikian, teknik pengambilan sampel berbasis pengacakan terbukti berguna dalam desain sistem pengenalan pola di lingkungan yang bermusuhan.

Biggio dkk. memperluas lingkungan adversarial dalam pengklasifikasi linier ke sistem pengklasifikasi ganda (MCS) berbasis pengacakan. MCS menggabungkan pengklasifikasi dasar linier melalui pengantongan dan pengambilan sampel subruang acak. Untuk meningkatkan akurasi dan ketahanan klasifikasi, distribusi bobot MCS diselidiki untuk mendapatkan distribusi nilai bobot yang lebih merata dibandingkan bobot pengklasifikasi tunggal. Dalam serangan terburuk, musuh diasumsikan memiliki pengetahuan lengkap tentang rangkaian fitur, parameter pengklasifikasi, dan fungsi keputusan.

Dalam serangan yang bukan kasus terburuk, musuh diasumsikan memiliki pengetahuan yang tidak lengkap tentang fungsi keputusan pengklasifikasi. Ketahanan pengklasifikasi kemudian dievaluasi sebagai fungsi kekuatan serangan, yang mewakili jumlah maksimum fitur yang dapat dimodifikasi oleh musuh. Dalam serangan yang bukan kasus terburuk, musuh memperkirakan bobot fitur dengan melebih-lebihkan atau meremehkan pentingnya fitur yang paling diskriminan dalam kinerja klasifikasi. Dalam kasus serangan terburuk, musuh seharusnya memodifikasi fitur untuk meminimalkan fungsi pengambilan keputusan dan memaksimalkan penurunan kinerja pengklasifikasi.

Biggio dkk. secara formal mengukur kekerasan penghindaran untuk sistem klasifikasi pola penargetan musuh secara umum dan arsitektur pengklasifikasi ansambel pada khususnya. Kekerasan penghindaran diperhitungkan dalam pemilihan fitur dan pilihan arsitektur pengklasifikasi. Ini didefinisikan sebagai nilai yang diharapkan dari jumlah minimum fitur yang akan dimodifikasi oleh musuh yang berusaha menghindari pengklasifikasi. Ini dihitung pada himpunan bagian terpisah dari bobot fitur diskriminan (iid yang dikondisikan kelas) yang diasumsikan terdistribusi secara merata di antara beberapa pengklasifikasi dengan fungsi keputusan yang sama. Parameter pengklasifikasi dipilih untuk meminimalkan biaya klasifikasi yang diberikan dalam bentuk kesalahan positif palsu dan negatif palsu.

3.5 MODEL PENGELOMPOKAN

Masalah pengelompokan adversarial tidak dapat diselesaikan dengan kriteria stabilitas pengelompokan yang mengatasi gangguan stokastik dalam kumpulan data, dibandingkan dengan manipulasi yang ditargetkan. Biggio dkk. merancang serangan keracunan dan kebingungan untuk pengelompokan hierarki hubungan tunggal. Dalam serangan keracunan seperti itu, tujuan pihak musuh adalah memasukkan contoh-contoh adversarial ke dalam ukuran kualitas pengelompokan. Dalam serangan kebingungan, tujuan musuh adalah menyembunyikan sampel data di cluster yang ada dengan memanipulasi nilai fitur. Dalam serangan ini, cluster didefinisikan tidak hanya sebagai partisi keras (dan partisi lunak) dari algoritma clustering partisi tetapi juga sebagai set dominan (dan hierarki subset yang diparameterisasi) dalam algoritma clustering tipe linkage. Biggio dkk. memberikan

batasan tambahan pada skenario serangan dengan metrik jarak antara data pelatihan yang tidak dimanipulasi dan data adversarial yang dimanipulasi.

Tingkat pengetahuan musuh dikodekan oleh entropi distribusi probabilitas dalam sampel serangan. Distribusi probabilitas ditentukan pada ruang pengetahuan musuh yang memberikan informasi tentang kumpulan data dan parameterisasinya dalam algoritma pengelompokan. Misalkan pengetahuan adversarial seperti itu, tujuan musuh adalah fungsi tujuan yang dinyatakan dalam bentuk (i) ukuran jarak bernilai nyata antara pengelompokan yang mengevaluasi sampel serangan untuk serangan keracunan dan (ii) ukuran divergensi skalar nyata non-negatif antara sampel serangan dan sampel target. Di sini, kriteria pemotongan dendrogram heuristik serakah mewakili keluaran pengelompokan hierarki tautan tunggal sebagai matriks biner probabilitas yang menugaskan sampel ke kluster.

3.6 MODEL PEMILIHAN FITUR

Xiao dkk. memberikan kerangka manipulasi untuk menyelidiki serangan keracunan pada metode pemilihan fitur seperti LASSO, regresi ridge, dan jaring elastis. Pemilihan fitur tersebut digunakan untuk memperoleh informasi sensitif keamanan yang dapat ditindaklanjuti dalam teknologi berbasis data berdimensi tinggi berskala besar seperti deteksi spam, deteksi malware, peringkat halaman web, dan verifikasi protokol jaringan. Xiao dkk. mengasumsikan pemilihan fitur sebagai masalah memfilter subset fitur yang relevan yang menyimpulkan proses acak iid untuk data pelatihan. Kriteria pemilihan fitur kemudian direpresentasikan sebagai optimasi fungsi tujuan seperti kesalahan klasifikasi dan perolehan informasi prediksi.

Xiao dkk. mendefinisikan tujuan musuh dalam kaitannya dengan pelanggaran keamanan yang dapat dikategorikan sebagai salah satu pelanggaran integritas, pelanggaran ketersediaan, dan pelanggaran privasi dalam pemilihan fitur. Pelanggaran integritas sedikit mengubah subset fitur yang dipilih untuk memfasilitasi serangan penghindaran berikutnya. Pelanggaran ketersediaan membahayakan algoritme pemilihan fitur untuk menghasilkan subset fitur keluaran dengan kesalahan generalisasi terbesar. Pelanggaran privasi merekayasa balik proses pemilihan fitur untuk menyimpulkan informasi tentang subset fitur, data pelatihan, dan pengguna sistem. Serangan yang ditargetkan memengaruhi subset fitur tertentu, sedangkan serangan sembarangan memengaruhi pemilihan fitur apa pun.

Xiao dkk. misalkan pengetahuan musuh dapat berupa asumsi pada data pelatihan, representasi fitur, algoritma pemilihan fitur, dan kriteria pemilihan fitur. Kemudian pengaruh musuh dapat bersifat kausatif atau eksploratif yang masing-masing memengaruhi data pelatihan atau data pengujian. Di sini, serangan keracunan untuk pemilihan fitur memanipulasi nilai dan label fitur dalam data pelatihan untuk membuat sampel keracunan yang selanjutnya akan salah diklasifikasikan. Serangan penghindaran untuk pemilihan fitur memanipulasi data pengujian untuk menghindari deteksi dengan mengusulkan pengukuran jarak dan strategi adversarial untuk membandingkan data asli, sampel serangan yang tidak dimanipulasi, dan sampel serangan. Strategi adversarial tersebut dinyatakan dalam bentuk pengetahuan musuh, kemampuan musuh, tujuan musuh, dan pengaruh musuh dalam

mempengaruhi penghitungan fungsi kerugian adversarial yang secara empiris menyimulasikan algoritma pemilihan fitur pada data yang diracuni.

Untuk setiap algoritma pemilihan fitur dalam percobaan, Xiao et al. mengoptimalkan fungsi kerugian adversarial dengan algoritma pendakian (sub)gradien yang memecahkan masalah optimasi cembung. Ruang fitur diasumsikan mendefinisikan fitur kontinu dan diskrit serta fitur terdiferensiasi dan tak terdiferensiasi. Untuk mengevaluasi pemilihan fitur dalam pengaturan serangan, indeks stabilitas diusulkan untuk menunjukkan peringkat anti-korelasi antara subset fitur dari algoritma pemilihan fitur. Eksperimen menunjukkan bahwa musuh dapat dengan mudah mengkompromikan algoritme pemilihan fitur yang mendorong ketersebaran dalam representasi fitur. Serangan peracunan dan serangan penghindaran dikatakan menyesatkan pengambilan keputusan model dengan memasukkan bias model dan varians model, masing-masing, ke dalam dekomposisi kesalahan kuadrat rata-rata algoritma pemilihan fitur.

Mei dkk. meracuni korpus alokasi Dirichlet (LDA) laten sehingga LDA menghasilkan topik yang dimanipulasi secara adversarial dalam keputusan pengguna LDA. Serangan adversarial dirumuskan sebagai masalah optimasi bilevel untuk inferensi variasional dalam batasan anggaran. Ini diselesaikan dengan metode penurunan gradien yang efisien secara komputasi berdasarkan fungsi implisit. Pengoptimalan ini menggunakan perbedaan KL antara distribusi topik kata pembelajar LDA dan distribusi variasional yang sepenuhnya terfaktor yang dibatasi oleh kondisi Karush-Kuhn-Tucker (KKT). Musuh meracuni korpus pelatihan sedemikian rupa sehingga topik yang dipelajari oleh LDA dipandu menuju distribusi multinomial target yang ditentukan oleh musuh. Tujuan Adversary adalah meminimalkan fungsi risiko penyerang yang menentukan jarak antara distribusi multinomial adversarial dan distribusi multinomial pelatihan. Risiko musuh dikombinasikan dengan divergensi KL pembelajar memberikan kerangka optimasi dua tingkat untuk membangun contoh adversarial. Contoh adversarial mengenai kata dan kalimat yang menyesatkan topik LDA dibuat pada korpus yang bersumber dari transkrip debat Dewan Perwakilan Rakyat Amerika Serikat, ucapan selamat tahun baru online, dan artikel kawat berita TREC AP.

3.7 MODEL DETEKSI ANOMALI

Kloft dkk. mengeksplorasi contoh adversarial untuk algoritma deteksi anomali (centroid online). Skenario serangan adversarial dinyatakan dalam efisiensi dan kendala dalam merumuskan serangan optimal pada deteksi outlier. Deteksi outlier menemukan peristiwa yang tidak biasa di seluruh jendela geser terbatas dalam aplikasi keamanan komputer seperti pembuatan tanda tangan otomatis dan sistem deteksi intrusi. Serangan keracunan diasumsikan menciptakan contoh adversarial pada data pelatihan di mana persentase tertentu dari data pelatihan dikendalikan oleh musuh. Titik data anomali kemudian diukur berdasarkan jarak Euclidean dari rata-rata empiris data pelatihan. Rata-rata empiris dihitung pada data pelatihan dengan algoritma online jendela geser terbatas untuk data non-stasioneritas. Dengan mendorong titik rata-rata empiris ke arah contoh yang

berlawanan, musuh memaksa algoritme deteksi anomali untuk menerima titik data anomali sebagai data pelatihan normal.

Kloft dkk. menyatakan perpindahan relatif rata-rata empiris asli dalam bentuk vektor arah serangan antara titik serangan dan titik rata-rata. Serangan optimal serakah kemudian diusulkan untuk menemukan titik serangan di sel Voronoi pada titik data yang memaksimalkan perpindahan relatif dari mean empiris. Untuk norma Euclidean, serangan serakah dioptimalkan dengan program linier atau program kuadrat. Pencampuran titik normal dan titik serang dimodelkan dengan variabel acak Bernoulli yang iid dalam ruang kernel Hilbert. Kemajuan serangan diukur dengan memproyeksikan mean empiris saat ini ke vektor arah serangan. Analisis teoretis diberikan untuk membatasi ekspektasi dan varians perpindahan relatif dengan jumlah titik latihan dan titik serangan pada jendela geser saat ini. Musuh diasumsikan memiliki pengetahuan penuh tentang data pelatihan dan algoritma deteksi anomali. Pertahanan detektor anomali terhadap serangan musuh diusulkan dalam hal mengendalikan tingkat positif palsu.

Rubinstein dkk. mengevaluasi serangan keracunan dan pertahanan pelatihan untuk deteksi anomali subruang analisis komponen utama (PCA) di mana komponen utama memaksimalkan ukuran penyebaran data pelatihan yang kuat. Sasaran musuh dinyatakan dengan meningkatnya positif palsu dan negatif palsu dari model yang diserang. Rangkaian waktu volume lalu lintas antara pasangan titik adalah kumpulan data yang mewakili matriks perutean. PCA yang kuat dari matriks perutean kemudian mengidentifikasi anomali volume di subruang abnormal. Strategi peracunan musuh mempertimbangkan serangan dengan jumlah informasi varians yang semakin meningkat dalam skenario serangan. Strategi serangan terlemah tidak mengetahui apa pun tentang arus lalu lintas dan menambahkan kebisingan acak sebagai contoh adversarial. Dalam strategi serangan yang diinformasikan secara lokal, musuh menyadap informasi tentang volume lalu lintas saat ini pada tautan jaringan yang diserang. Dalam strategi serangan yang diinformasikan secara global, musuh memiliki pengetahuan tentang volume lalu lintas di semua link jaringan dan tingkat jaringan.

Dalam serangan jangka pendek, detektor anomali dilatih ulang untuk setiap minggu data pelatihan selama musuh menyerang jaringan. Dalam serangan jangka panjang, komponen utama detektor anomali secara perlahan diracuni oleh musuh selama beberapa minggu. Dalam setiap skenario serangan, musuh memutuskan jumlah data yang akan ditambahkan ke arus lalu lintas target berdasarkan variabel acak Bernoulli. Analisis PCA yang kuat pada matriks perutean yang diubah secara berlawanan kemudian menghasilkan contoh-contoh adversarial yang diklasifikasikan sebagai tidak berbahaya oleh detektor anomali. Solusi analitik yang dapat diterapkan untuk PCA yang kuat diperoleh oleh Rubinstein et al. dari fungsi tujuan dengan perkiraan relaksasi yang memaksimalkan vektor serangan yang diproyeksikan ke matriks kovarians subruang normal. Metode pengejaran proyeksi kemudian menghasilkan solusi layak untuk fungsi tujuan dalam arah gradiennya.

Feng dkk. menyajikan outlier yang merugikan untuk regresi logistik. Prosedur pemrograman linier memperkirakan parameter logistik dengan adanya adversarial outlier dalam matriks kovarians dalam masalah klasifikasi biner. Ketidaktahanan regresi logistik

terhadap adversarial outlier dihitung dari estimasi kemungkinan maksimum fungsi pengaruh log-likelihood serta fungsi kerugian dalam data pelatihan dimensi tinggi yang telah dirusak oleh adversarial outlier. Dalam skenario serangan, adversarial outlier berusaha mendominasi korelasi dalam fungsi tujuan model regresi logistik. Batasan ketahanan kemudian diturunkan pada risiko populasi dan risiko empiris dengan fungsi kerugian berkelanjutan Lipschitz.

3.8 MODEL HUBUNGAN MULTITASKING

Zhao dkk. mengusulkan serangan peracunan data pada keterkaitan tugas dalam pembelajaran hubungan multi-tugas (MTRL). Serangan optimal di MTRL memecahkan masalah optimasi bilevel yang adaptif terhadap tugas target yang berubah-ubah dan tugas menyerang. Serangan tersebut ditemukan dengan prosedur pendakian gradien stokastik. Kerentanan MTRL terhadap contoh adversarial dikategorikan menjadi pendekatan pembelajaran fitur, pendekatan peringkat rendah, pendekatan pengelompokan tugas, dan pendekatan hubungan tugas yang tujuan pembelajarannya adalah untuk bersama-sama mempelajari suatu fungsi prediksi. Kemudian dipelajari MTRL fungsi prediksi linier dengan fungsi kerugian cembung sembarang dan matriks kovarian semi pasti positif.

Sasaran Musuh didefinisikan sebagai penurunan kinerja serangkaian tugas target dengan memasukkan data beracun ke serangkaian tugas penyerang. Fungsi pembayaran musuh didefinisikan sebagai hilangnya data pelatihan empiris pada tugas target di mana musuh memiliki pengetahuan lengkap tentang model MTRL target. Dalam prosedur pendakian gradien, data keracunan diperbarui secara berulang ke arah memaksimalkan fungsi hasil adversarial. Fungsi prediksi adalah fungsi kerugian kuadrat terkecil untuk tugas regresi dan fungsi kerugian engsel kuadrat untuk tugas klasifikasi. Kinerja prediksi dievaluasi dengan memaksimalkan area di bawah kurva untuk tugas klasifikasi dan meminimalkan kesalahan kuadrat rata-rata yang dinormalisasi untuk tugas regresi.

3.9 MODEL REGRESI

Liu dkk. mempelajari pembelajaran yang diawasi adversarial dalam masalah regresi dimensi tinggi. Sasaran manipulasi adversarial adalah data pelatihan. Skenario serangannya disebut serangan keracunan. Berbeda dengan model pembelajaran terawasi yang kuat yang membuat asumsi statistik yang kuat tentang distribusi masukan yang mendasari dan sifat selanjutnya dari matriks fitur, independensi fitur, dan rasio sinyal-kebisingan, usulan pembelajaran terawasi adversarial melonggarkan asumsi tersebut untuk memperkirakan matriks fitur dengan tingkat yang rendah. matriks peringkat yang cocok untuk regresi yang kuat. Jaminan kinerja yang dihasilkan dibandingkan dengan regresi komponen utama yang kuat yang bertindak sebagai model dasar.

Model pembelajaran mesin tersebut dapat diterapkan dalam pemfilteran spam, analisis lalu lintas, dan deteksi penipuan untuk menegakkan keamanan terhadap musuh yang kuat. Tantangan pembelajaran yang harus diatasi dalam desain algoritma adalah bahwa pengurangan dimensi dapat memulihkan pola subruang peringkat rendah secara andal dan regresi yang dilakukan pada subruang dapat memulihkan prediksi yang akurat. Lebih lanjut,

tujuan desain ini harus dicapai meskipun ada sampel yang diracuni dalam kumpulan data pelatihan. Untuk mencapai tujuan ini, penulis mengembangkan algoritma faktorisasi matriks yang kuat yang memulihkan subruang dengan benar jika memungkinkan dan menggunakan fitur-fiturnya dalam regresi komponen utama yang dipangkas, yang menggunakan basis pemulihan dan optimasi yang dipangkas untuk memperkirakan parameter model linier.

Sisa kebisingan adalah solusi regresi yang kuat. Ini digunakan untuk mempelajari interferensi data adversarial dengan desain model regresi dan kemampuan adversarial untuk mendistorsi estimator secara signifikan. Hal ini mengarah pada desain fungsi kerugian terbatas untuk pembelajaran adversarial. Musuh kemudian dapat diasumsikan menciptakan strategi keracunan untuk memicu kinerja kasus terburuk dalam algoritma reduksi dimensi dan model regresi. Serangan yang paling efektif adalah memindahkan sampel data ke sepanjang arah untuk memodifikasi estimator yang dipelajari secara maksimal. Hasil eksperimen dibandingkan dengan model regresi linier yang dirancang agar tahan terhadap peracunan data yang merugikan.

Model pembelajaran adversarial seperti itu cenderung lebih fokus pada pertahanan terhadap musuh untuk menghasilkan algoritme pembelajaran adversarial dengan ketahanan distribusi daripada menyiapkan skenario serangan untuk memvalidasi biaya kesalahan kesalahan klasifikasi. Prediksi pemodelan regresi non-linier selanjutnya dapat memperoleh manfaat dari rekayasa fitur keamanan berdasarkan teori pembelajaran adversarial yang melibatkan model pembelajaran representasi mendalam seperti mesin faktorisasi. Mesin faktorisasi adalah perkiraan peringkat rendah dari tensor data renggang ketika sebagian besar elemen prediksinya tidak diketahui. Dalam rekayasa fitur keamanan, mereka dapat memodelkan interaksi antar fitur menggunakan parameter yang difaktorkan. Mereka berlaku tidak hanya untuk tugas reduksi dimensi tetapi juga untuk tugas prediksi umum dalam pengaturan dimensi tinggi.

Seperti yang dikemukakan oleh Blondel dkk., mesin faktorisasi tingkat tinggi dapat diperkirakan dengan algoritma pemrograman dinamis yang disesuaikan untuk tugas prediksi dalam pembelajaran adversarial. Aplikasi dapat didemonstrasikan untuk aplikasi prediksi tautan dalam jaringan yang kompleks. Dalam pembelajaran adversarial teoretis permainan, algoritma pemrograman dinamis dapat diusulkan untuk mempelajari sifat konvergensi optima teoretis permainan. Fungsi kerugian teoretis permainan dan prosedur pelatihan dalam penelitian semacam itu dapat diterapkan pada studi pembelajaran dan pengambilan sampel ulang dinamika dalam mekanisme komputasi saraf yang disesuaikan dengan pembelajaran adversarial. Dinamika kompleks yang diekspresikan dalam distribusi data yang berlawanan kemudian dapat dimodelkan sebagai algoritma pengacakan dalam sistem penambangan data dan model pembelajaran mesin.

Amin dkk. mempelajari regresi adversarial dalam sistem cyber-fisik (CPS). CPS merupakan infrastruktur penting seperti jaringan tenaga listrik, jaringan transportasi, jaringan air, dan pembangkit listrik tenaga nuklir. Model regresi terawasi diusulkan untuk mendeteksi pembacaan sensor yang tidak wajar di infrastruktur tersebut. Kemudian model teori permainan dibangun berdasarkan interaksi antara pembela CPS dan musuh. Di dalamnya,

pembela memilih ambang batas deteksi, sementara penyerang melancarkan serangan diam-diam sebagai respons. Serangan semacam ini disebabkan oleh modifikasi hati-hati terhadap pembacaan sensor yang disusupi agar tidak terdeteksi. Serangan Stuxnet diberikan sebagai contoh terkenal yang menargetkan infrastruktur fisik melalui cara siber.

Hal ini didefinisikan sebagai kerusakan pembacaan sensor untuk memastikan bahwa serangan terhadap kode pengontrol tidak terdeteksi atau secara tidak langsung berdampak pada perilaku pengontrol. Masalah pembelajaran memecahkan deteksi anomali adversarial dalam konteks serangan integritas pada subset sensor di CPS. Tugas regresi yang diawasi dalam model deteksi anomali memprediksi pengukuran untuk setiap sensor sebagai fungsi pembacaan dari sensor lainnya. Deteksi anomali yang kuat dimodelkan sebagai permainan Stackelberg antara pemain bertahan dan musuh. Prediktor ansambel yang berisi kombinasi regresi jaringan saraf dan regresi linier dieksplorasi dalam detektor berbasis regresi. Tujuan adversarial dinyatakan sebagai masalah program linier bilangan bulat campuran. Dengan demikian, fungsi kerugian adversarial dapat diturunkan untuk masalah pengambilan sampel, prediksi, dan pengoptimalan dalam regresi pembelajaran mendalam. Garis dasar regresi tersebut dapat dibangun untuk masalah prediksi multivariat dalam pembelajaran adversarial.

Permainan Stackelberg dapat digunakan untuk memodelkan interaksi strategis dengan asumsi agen rasional di pasar di mana terdapat persaingan hierarkis. Interaksi strategis antara fungsi pembayaran untuk kedua pemain mencerminkan peringkat relatif skenario aplikasi masing-masing pemain dalam hal hasil akhir yang diharapkan dalam pembelajaran mesin. Ruang pencarian strategi untuk setiap pemain dalam permainan biasanya diasumsikan berbatas dan cembung, dan fungsi pembayaran yang sesuai diasumsikan dapat dibedakan. Solusi ekuilibrium untuk semua fungsi pembayaran dalam permainan ditentukan oleh solusi fungsi tujuan optimasi. Teori permainan menyediakan alat matematika untuk memodelkan perilaku pembela dan perilaku musuh dalam pembelajaran mesin dalam hal strategi pertahanan dan serangan. Pembelajaran adversarial teoritis permainan memperhitungkan trade-off yang dilakukan oleh penyerang antara biaya beradaptasi dengan pengklasifikasi dan manfaat dari serangan. Di sisi lain, pengorbanan yang dilakukan oleh pembela HAM menyeimbangkan antara manfaat deteksi serangan yang benar dan kerugian jika terjadi alarm palsu.

Zhang dkk. mempertimbangkan model regresi yang kuat di lingkungan online dan terdistribusi. Di sini, aplikasi penambangan data dari teori pembelajaran adversarial harus mengakomodasi tantangan baru pada metodologi analisis data yang diterapkan oleh data besar, di mana biasanya tidak mungkin untuk menyimpan seluruh aliran data berdimensi tinggi atau memindainya berkali-kali karena volumenya yang sangat besar. dan mengubah dinamika distribusi data yang mendasarinya dari waktu ke waktu. Kita harus mempertimbangkan biaya komputasi yang diamortisasi dari penambangan pola pada distribusi data berdimensi tinggi dan multidimensi. Metrik validasi pola dalam pembelajaran adversarial juga harus mempertimbangkan pengetahuan domain fisik sebagai kebenaran dasar untuk pembelajaran yang diawasi.

Pemodelan ini perlu mempelajari substruktur padat, kelas langka, dan pola padat pada kumpulan data transaksional, sekuensial, dan grafik di mana proses acak yang menghasilkan data pelatihan mungkin tidak sama dengan proses yang mengatur data pengujian. Pendekatan yang diterapkan harus digabungkan dengan pendekatan pembelajaran diskriminatif adaptif dengan optimalisasi berkelanjutan. Di sini, contoh-contoh adversarial dapat dimasukkan ke dalam skenario serangan yang jarang, terukur, dan padat untuk pembelajaran adversarial. Mereka harus memperhitungkan aliran analisis data operasional yang sensitif terhadap latensi dan pembelajaran representasi pengetahuan melalui sumber daya terbatas yang ditentukan dalam hal waktu, daya, dan biaya komunikasi. Manipulasi adversarial harus didefinisikan melalui pembelajaran bertahap dan pemrosesan serangan adversarial yang terdistribusi pada kecepatan dan skala data yang besar.

Zhang dkk. mengidentifikasi kerusakan data yang terdistribusi secara heterogen karena musuh. Mereka mempunyai usulan untuk memperkirakan kerusakan ketika data tidak dapat seluruhnya dimuat ke dalam memori komputer. Regresi yang kuat dilakukan dengan model regresi kuadrat terkecil yang dapat diskalakan yang mempelajari sekumpulan koefisien regresi yang andal. Algoritme online dan terdistribusi diusulkan untuk regresi yang kuat tersebut. Koefisien regresi yang sebenarnya dipulihkan dengan batas atas yang konstan pada kesalahan metode batch yang canggih di bawah asumsi korupsi sewenang-wenang yang tidak terdistribusi secara merata dalam batch mini data pelatihan yang disediakan untuk algoritma online.

3.10 MANIPULASI PEMBELAJARAN MESIN DALAM KEAMANAN SIBER

Serangan adversarial memiliki beberapa penerapan dalam visi komputer, pemrosesan bahasa alami, keamanan dunia maya, dan dunia fisik. Dalam visi komputer, serangan adversarial diciptakan untuk klasifikasi gambar dan deteksi objek. Dalam pemrosesan bahasa alami, serangan adversarial diciptakan untuk klasifikasi teks dan terjemahan mesin. Dalam keamanan dunia maya, serangan adversarial diciptakan untuk layanan cloud, deteksi malware, dan deteksi intrusi.

Di dunia fisik, serangan adversarial diciptakan untuk menskalakan pelatihan adversarial ke model dan kumpulan data besar. Kurakin dkk. mendiskusikan skenario dunia fisik dengan kamera dan sensor lain sebagai masukan. Eykholt dkk. menghasilkan gangguan visual yang kuat dalam kondisi fisik yang berbeda untuk kasus klasifikasi rambu jalan di dunia nyata. Di sini, tugas visi komputer bertindak sebagai jalur kontrol dalam sistem fisik di mana tantangan utama dalam menghasilkan gangguan fisik yang kuat adalah variabilitas lingkungan. Melis dkk. membuat serangan dunia fisik pada sistem penglihatan robot. Sharif dkk. membuat serangan dunia fisik pada sistem biometrik wajah menggunakan model pengenalan wajah untuk pengawasan dan kontrol akses.

Xiao dkk. membahas transformasi spasial dari gangguan adversarial dengan jarak Lp yang bertindak sebagai metrik kualitas persepsi dalam menghukum gangguan adversarial. Akhtar dkk. mensurvei serangan adversarial terhadap pembelajaran mendalam dalam visi komputer. Dibandingkan dengan serangan teoretis dalam game, serangan dunia fisik secara

fisik mengubah tampilan suatu objek untuk menipu deteksi terlatih. Mereka dibatasi hanya untuk permainan serangan sekali saja yang diterapkan pada ancaman yang ditargetkan dalam pengaturan teoritis permainan. Mereka tampaknya dapat diterapkan untuk menghasilkan kebijakan pencarian stokastik dalam permainan kami dengan heuristik pencarian seperti pencarian pohon Monte Carlo dalam teori permainan kombinatorial.

Metode pembelajaran mendalam dapat digunakan untuk memajukan tujuan keamanan siber seperti deteksi, pemodelan, pemantauan, analisis, dan pertahanan terhadap berbagai ancaman terhadap data sensitif dan sistem keamanan. Rossler dkk. membahas pembuatan gambar sintetis dan tolok ukur manipulasi berdasarkan DeepFakes, Face2Face, FaceSwap, dan NeuralTextures sebagai perwakilan terkemuka untuk manipulasi wajah dalam detektor pemalsuan berbasis data. Matern dkk. memamerkan artefak dari pelacakan dan pengeditan wajah untuk mengungkap manipulasi dalam algoritma pengeditan wajah seperti DeepFakes dan Face2Face.

Penerapan lebih lanjut dari pembelajaran mesin adversarial dalam keamanan siber mencakup deteksi malware, klasifikasi malware, deteksi spam, deteksi phishing, deteksi botnet, deteksi intrusi dan pencegahan intrusi, serta deteksi anomali. Tong dkk. membahas serangan penghindaran dalam deteksi malware PDF. Melis dkk. menerapkan model pembelajaran mesin yang dapat dijelaskan dalam deteksi malware Android. Marino dkk. menjelaskan klasifikasi yang salah dalam sistem deteksi intrusi berbasis data. Korona dkk. memberikan taksonomi serangan adversarial dalam sistem deteksi intrusi (IDS) dan infrastruktur komputasi. Demetrio dkk. mengusulkan atribusi fitur untuk memberikan penjelasan yang berarti terhadap klasifikasi biner malware. Fleshman dkk. mengukur ketahanan sistem produk anti-virus berbasis pembelajaran mesin menggunakan model deteksi malware.

Analisis Sensitivitas Manipulasi Pembelajaran Mesin Deep Learning

Memahami model pembelajaran mesin berguna untuk memvalidasi kebenarannya, mendeteksi bias algoritmik dan kebocoran informasi, serta mempelajari pola baru dari data. Khususnya dalam model pembelajaran mesin yang kompleks, performa canggih untuk model tersebut harus dibayar dengan kemampuan interpretasi. Kita harus menyeimbangkan antara kemampuan belajar dan ketahanan pembelajaran mesin yang diawasi. Yang kami maksud dengan “kemampuan belajar” adalah kemampuan pengklasifikasi untuk memprediksi label yang benar (tanpa memperhatikan noise), dan dengan “kekokohan”, yang kami maksudkan adalah bahwa prediksi tersebut akan sama dengan atau tanpa noise (tanpa memperhatikan kebenarannya). Kerugian yang kami amati adalah bahwa semakin banyak kemampuan untuk dipelajari akan mengakibatkan semakin berkurangnya ketahanan, dan sebaliknya.

Analisis sensitivitas adalah studi tentang pengaruh suatu variabel terikat terhadap perubahan variabel bebas. Hal ini berguna dalam mempelajari skenario serangan kotak hitam dalam pembelajaran adversarial di mana keluaran model dan proses pembelajaran merupakan fungsi buram dari beberapa masukan. Artinya, hubungan pasti antara masukan dan keluaran untuk pembelajaran mesin tidak dipahami dengan baik secara analitis. Analisis sensitivitas dalam pembelajaran mesin memiliki peran penting dalam analisis sistem

kompleks untuk kecerdasan buatan. Hal ini dapat digunakan untuk menentukan model dari sistem yang diteliti. Ini dapat mengidentifikasi parameter model yang berkontribusi pada faktor variabilitas keluaran analisis data.

Hal ini dapat mengidentifikasi wilayah pencarian optimal yang diminati dalam studi kalibrasi pada faktor analitik dan interaksinya. Terakhir, hal ini dapat mengevaluasi model analitik untuk membuat distribusi output dari respons berpengaruh yang dinilai dalam metode analitik untuk korelasi/klasifikasi/regresi, inferensi Bayesian, dan pembelajaran mesin. Tinjauan mengenai ukuran sensitivitas seperti dalam analisis varians diberikan oleh Frey dkk. Sebagai metode penilaian risiko dalam representasi pembelajaran adversarial, ukuran sensitivitas dapat digunakan untuk memprioritaskan pengumpulan data tambahan, mengidentifikasi titik kontrol penting dalam kumpulan data, dan memverifikasi validasi model. Analisis sensitivitas dapat dilakukan untuk penemuan pengetahuan, pemeringkatan fitur, reduksi dimensi, dan penyetelan model.

Berbeda dengan analisis sensitivitas, analisis skenario memeriksa skenario aplikasi spesifik untuk pembelajaran mesin dengan sangat detail untuk menemukan semua variabel relevan yang selaras dengan skenario tersebut. Variabel-variabel ini pada gilirannya akan mendukung penciptaan basis pengetahuan yang memahami keseluruhan hasil dengan serangkaian variabel masukan tertentu yang menentukan skenario dunia nyata. Dengan menguji model pembelajaran mesin di berbagai skenario, analisis sensitivitas menambah kredibilitas model tersebut dengan menginformasikan pengambilan keputusan berdasarkan data menuju kesimpulan nyata dan keputusan optimal. Perhatian khusus dapat diberikan pada bias algoritmik yang mendukung fitur langka dengan sensitivitas kuat terhadap kesalahan estimasi probabilitas dan proses gangguan yang merugikan. Pelatihan adversarial mengacu pada penggabungan contoh adversarial ke dalam proses pelatihan model pembelajaran mesin.

Pelatihan adversarial tidak hanya sensitif terhadap parameter tetapi juga hyperparameter yang memengaruhi proses pelatihan. Duesterwald dkk. menyajikan analisis sensitivitas lanskap hyperparameter dalam pelatihan adversarial jaringan pembelajaran mendalam. Teknik pengoptimalan hyperparameter diterapkan untuk menyesuaikan pelatihan adversarial guna memaksimalkan ketahanan sekaligus menjaga hilangnya akurasi dalam anggaran yang ditentukan. Wexler dkk. mengembangkan Alat Bagaimana-Jika yang memungkinkan analisis sistem pembelajaran mesin untuk menyelidiki dan memvisualisasikan masukan dan keluarannya. Ini dapat digunakan untuk menganalisis kepentingan fitur, menguji kinerja dalam pengujian hipotesis, dan memvisualisasikan perilaku model di beberapa kumpulan data masukan.

Alat seperti itu menarik bagi praktisi pembelajaran mesin untuk menjawab pertanyaan tentang pengaruh manipulasi adversarial terhadap titik data pada prediksi pemodelan. Ini juga dapat digunakan untuk menganalisis ketahanan distribusi model pembelajaran mesin di seluruh sampel data yang bertindak sebagai kumpulan data pelatihan, pengujian, dan validasi. Pengguna Alat Bagaimana-Jika memiliki antarmuka visual untuk melakukan penalaran kontrafaktual, menyelidiki batasan keputusan, dan mengeksplorasi perubahan prediksi

sehubungan dengan perubahan titik data. Oleh karena itu, hal ini mendukung pembuatan prototipe dan eksplorasi cepat terhadap berbagai hipotesis statistik dalam pembelajaran adversarial.

Tanpa akses ke detail pemodelan, penjelasan yang dapat digeneralisasikan dapat dihasilkan untuk manipulasi adversarial menggunakan hipotesis tersebut dengan cara model-agnostik. Fleksibilitas penjelasan dan representasinya meningkatkan kemampuan interpretasi pembelajaran adversarial. Hal ini dapat dikombinasikan dengan proses analisis data eksplorasi untuk menangani kompleksitas dalam tipe data masukan, tugas pemodelan, dan strategi pengoptimalan. Alur kerja untuk menguji skenario hipotetis di Alat Bagaimana-Jika mendukung pemahaman umum seputar data selain evaluasi metrik kinerja yang dioptimalkan menuju batasan keadilan pada pembelajaran mesin. Titik data dalam prediksi keluaran dapat divisualisasikan dengan kebingungan matriks dan kurva ROC.

Dalam pembelajaran mesin yang diawasi, analisis sensitivitas mempelajari kemungkinan kesalahan klasifikasi karena gangguan bobot dalam model pembelajaran yang disebabkan oleh ketidakakuratan mesin dan masukan yang berisik. Untuk memvalidasi ketahanan distribusi pembelajaran yang diawasi pada berbagai masukan, analisis sensitivitas telah diperluas ke dalam teknik pengoptimalan untuk jaringan saraf seperti reduksi sampel, pemilihan fitur, pembelajaran aktif, dan pembelajaran adversarial. Setelah membahas pendekatan geometris dan statistik terhadap analisis sensitivitas pembelajaran mesin, Yeung dkk. memamerkan penerapannya dalam pengurangan dimensi, optimalisasi jaringan, dan pembelajaran selektif.

Asalkan jaringan saraf mengandung jumlah unit tersembunyi yang optimal dan mampu membangun batas pembedaan yang optimal antar kelas, jaringan tersebut dapat digunakan untuk ekstraksi fitur dan induksi aturan dengan analisis sensitivitas pada pola informatif yang dinyatakan dalam batas keputusan. Engelbrecht dkk. mengusulkan analisis sensitivitas pada batas keputusan jaringan saraf dan menyajikan algoritma visualisasinya. Pola dinamis yang ditemukan dari analisis sensitivitas digunakan dalam algoritma pembelajaran selektif. Dengan demikian, kita dapat mengekstrak aturan yang akurat dari jaringan saraf terlatih. Patwary dkk. melakukan investigasi pembelajaran semi-supervisi berbasis analisis sensitivitas. Strategi membagi-dan-menaklukkan berdasarkan ketidakjelasan dalam kumpulan data pelatihan terbukti meningkatkan kinerja pengklasifikasi.

Di sini, pengklasifikasi mengklasifikasikan sebuah instance ke dalam kelas dengan tingkat keyakinan mengenai sejauh mana instance tersebut termasuk dalam kelas tertentu. Pada langkah pelatihan awal, pengklasifikasi dilatih pada sejumlah kecil data pelatihan dengan label kelas. Pada langkah pelatihan terakhir, sejumlah besar data tak berlabel digunakan untuk menetapkan setiap titik data ke salah satu dari beberapa label kelas. Kemampuan generalisasi pengklasifikasi pada kumpulan data validasi yang tidak terlihat terkait dengan ketidakjelasan pengklasifikasi dalam mencapai akurasi prediksinya. Sampel dengan ketidakjelasan rendah dari kumpulan data pengujian ditambahkan ke kumpulan data pelatihan asli untuk melatih ulang model pembelajaran dengan akurasi yang lebih baik. Metode resampling digunakan untuk mempelajari batas kesalahan generalisasi. Teori

pembelajaran dari data berisik seperti itu dapat digunakan untuk membangun pengklasifikasi pembelajaran semi-supervisi yang melibatkan metode pembelajaran seperti pelatihan mandiri, pelatihan bersama, pembelajaran multiview, maksimalisasi ekspektasi dengan model campuran generatif, penambangan pola grafik, dan SVM transduktif.

Suresh dkk. mengusulkan fungsi kerugian yang sensitif terhadap risiko untuk memecahkan masalah klasifikasi multi-kategori yang meminimalkan kesalahan perkiraan dan estimasi. Di sini, kesalahan perkiraan bergantung pada kedekatan prediksi dengan pengklasifikasi sebenarnya, dan kesalahan estimasi bergantung pada kedekatan distribusi masukan yang diperkirakan dengan distribusi masukan yang mendasarinya. Analisis kesalahan menggabungkan fungsi kerugian yang sensitif terhadap risiko ke dalam arsitektur pengklasifikasi jaringan saraf. Performa dibandingkan pada sampel pelatihan yang tidak seimbang menggunakan fungsi kerugian terkenal lainnya untuk memperkirakan probabilitas posterior.

Fungsi kerugian yang diusulkan meningkatkan akurasi klasifikasi secara keseluruhan dan per kelas. Fungsi kerugian yang sensitif terhadap risiko pada kinerja keputusan diperlukan untuk memperluas hasil klasifikasi biner ke masalah klasifikasi multi-kategori. Dalam jaringan saraf, pengklasifikasi menggunakan fungsi kerugian yang meminimalkan kesalahan klasifikasi yang diharapkan untuk semua kelas. Di sini, fungsi kerugian yang sensitif terhadap risiko mengukur tingkat kepercayaan dalam prediksi label kelas dan risiko terkait yang terkait dengan tindakan di balik setiap keputusan pengklasifikasi. Biaya kesalahan klasifikasi ditetapkan secara apriori. Dalam pemodelan teoritis permainan, kesalahan klasifikasi biaya seperti itu dapat dimasukkan ke dalam desain fungsi pembayaran yang merugikan. Dalam masalah klasifikasi multi-kategori, fungsi adversarial payoff harus mampu mengatasi tumpang tindih yang kuat antar kelas dalam data yang jarang dan ketidakseimbangan yang tinggi dalam sampel per kelas.

Arsitektur jaringan saraf kemudian menemukan distribusi probabilitas gabungan pada data observasi untuk mendapatkan estimasi akurat dari label kelas berkode yang diinginkan. Faktor risiko pada estimasi probabilitas posterior dalam desain fungsi kerugian memberikan dampak buruk pada pola kesalahan klasifikasi dan biayanya. Proses pelatihan pengklasifikasi dipandu oleh matriks konfusi dan matriks risiko. Kesulitan dalam memperoleh kebenaran dasar tentang kelas asli dalam distribusi data masukan meningkatkan kompleksitas komputasi dalam pengembangan model.

Cortez dkk. mengusulkan visualisasi untuk mengekstraksi pengetahuan yang dapat dipahami manusia dari model penambangan data kotak hitam pembelajaran yang diawasi menggunakan analisis sensitivitas. Model data mining yang dipertimbangkan adalah jaringan saraf, mesin vektor pendukung, hutan acak, dan pohon keputusan. Respon sensitivitas digunakan untuk membuat ukuran masukan yang penting untuk tugas regresi dan klasifikasi dalam penambangan data. Tren regresi yang ditambah sensitivitas dan pola klasifikasi yang ditemukan oleh data mining dapat digunakan untuk meningkatkan pengambilan keputusan berbasis data di dunia nyata. Di sini, model analitik berbasis data mempelajari fungsi dasar yang tidak diketahui yang memetakan beberapa variabel masukan ke satu target keluaran

dalam paradigma pembelajaran yang diawasi. Interpretabilitas model data mining dapat ditingkatkan dengan strategi rekayasa fitur seperti ekstraksi aturan dan teknik visualisasi multidimensi.

Analisis sensitivitas yang diusulkan memperlakukan model pembelajaran mesin sebagai kotak hitam untuk mengkueri model tersebut dengan sampel sensitivitas dan mencatat respons yang diperoleh. Itu tidak menggunakan informasi tambahan seperti kriteria kesesuaian model dan atribusi pentingnya fitur. Dasar pemikiran analisis sensitivitas adalah bahwa masukan yang relevan harus menghasilkan perubahan keluaran yang besar ketika tingkat masukannya diubah-ubah. Relevansi masukan tersebut diukur dengan menggunakan ukuran sensitivitas. Vektor dasar diusulkan untuk menangkap interaksi masukan dengan upaya komputasi yang lebih sedikit. Ukuran sensitivitas memodelkan hasil target dengan label kelas keluaran atau probabilitas kelas. Perhitungan area total di bawah kurva karakteristik operasi penerima (AUC) dan metode pengambilan sampel ansambel digunakan sebagai ukuran sensitivitas dalam tugas klasifikasi multi-label. Kesalahan regresi seperti mean absolute error (MAE) digunakan sebagai ukuran sensitivitas dalam tugas regresi multivariat.

Pengukuran sensitivitas pertama-tama dihitung untuk masing-masing kelas, dan kemudian rata-rata tertimbang dilakukan untuk menghitung pengukuran sensitivitas global. Efek dari ukuran sensitivitas ini pada peringkat pohon dan pemisahan hyperplane kemudian dipelajari dalam visualisasi data untuk berbagai masukan dalam eksperimen validasi silang yang memperkirakan metrik kinerja penambangan data. Beberapa metode analisis sensitivitas baru, ukuran, fungsi agregasi, dan teknik visualisasi diusulkan. Metode pemilihan fitur dapat dirancang lebih lanjut untuk meningkatkan peringkat relatif dalam ukuran sensitivitas untuk memandu pencarian melalui variabel yang relevan untuk tugas penambangan data yang ada.

Engelbrecht melakukan analisis sensitivitas batas keputusan yang dipelajari oleh fungsi keluaran jaringan saraf sehubungan dengan gangguan masukan. Analisis sensitivitas setiap fungsi aktivasi unit tersembunyi mengungkapkan batas mana yang diterapkan oleh unit tersembunyi mana. Di sini, batas keputusan diperlakukan sebagai wilayah ketidakpastian dalam klasifikasi ruang fitur masukan. Batasan keputusan diskriminatif yang unik dapat diperoleh dengan memangkas jaringan berukuran besar yang melebihi data, dengan mengembangkan jaringan berukuran kecil yang tidak memenuhi data, atau dengan menambahkan istilah regularisasi ke fungsi tujuan pembelajaran mesin.

Batasan keputusan yang optimal menghasilkan kinerja generalisasi pengklasifikasi yang baik secara statistik yang dihasilkan dengan mengacu pada persamaan batas keputusan. Metode pengambilan sampel data pada keluaran pemodelan digunakan untuk menemukan batasan keputusan dalam kaitannya dengan masukan pemodelan untuk pengklasifikasi. Parameter pemodelan yang tidak menentukan batasan keputusan apa pun dapat dipangkas karena tidak berkontribusi pada fungsi klasifikasi. Dengan memilih pola di wilayah melintasi batas keputusan, kita dapat melakukan pengurangan numerositas dan dimensi dalam kumpulan data pelatihan.

Fawcett mensurvei karakteristik grafik karakteristik operasi penerima (ROC) dalam tutorial. Kurva ROC dapat digunakan untuk mengatur keluaran pengklasifikasi dan memvisualisasikan kinerjanya. Mereka berakar pada teori deteksi sinyal tetapi telah diadopsi oleh komunitas pembelajaran mesin untuk menganalisis distribusi kelas yang miring, biaya kesalahan klasifikasi, dan pembelajaran yang sensitif terhadap biaya jika terdapat kelas yang tidak seimbang. Grafik ROC menggambarkan trade-off relatif antara manfaat (tingkat positif sebenarnya) dan biaya (tingkat positif palsu). Karena jaringan saraf dapat dianggap sebagai pengklasifikasi probabilistik yang menghasilkan probabilitas contoh relatif atau skor keanggotaan kelas, jaringan saraf tersebut dapat diberi peringkat dengan ambang batas untuk menghasilkan pengklasifikasi diskrit yang dapat menelusuri kurva melalui ruang ROC untuk membedakan antara contoh positif dan negatif.

Kombinasi penilaian dan pemungutan suara dapat dilakukan untuk menambah pengklasifikasi terpisah untuk menghasilkan tidak hanya label kelas tetapi juga perkiraan probabilitas yang diperlukan untuk membuat titik data dalam kurva ROC. Meskipun metrik kinerja pembelajaran mesin yang diperoleh dari matriks konfusi sensitif terhadap perubahan kelas untuk mencatat proporsi dalam data pelatihan, kurva ROC tidak sensitif terhadap perubahan distribusi kelas. Area di bawah kurva ROC (AUC) setara dengan probabilitas bahwa pengklasifikasi akan memberi peringkat pada instance positif yang dipilih secara acak lebih tinggi daripada instance negatif yang dipilih secara acak. Hal ini erat kaitannya dengan uji peringkat Wilcoxon dan indeks Gini. AUC dapat digunakan untuk membandingkan beberapa garis dasar klasifikasi.

Ada kemungkinan pengklasifikasi AUC tinggi memiliki kinerja lebih buruk di wilayah ruang ROC tertentu dibandingkan pengklasifikasi AUC rendah. Rata-rata kurva ROC memungkinkan kita memilih pengklasifikasi terbaik untuk sampel data pelatihan tertentu. Dengan menggunakan pohon estimasi probabilitas, AUC kelas jamak dapat dihasilkan untuk mempelajari diskriminasi antara beberapa pasangan kelas yang menggabungkan nilai diskriminabilitas berpasangan. Mengilustrasikan wilayah pembatas pengklasifikasi baru dalam grafik ROC dapat digunakan untuk mengarahkan induksi aturan dan konstruksi fitur dalam domain aplikasi untuk penambangan data.

Rekayasa fitur seperti itu dapat menghasilkan penemuan pola dalam hal model klasifikasi, aturan asosiasi, pola sekuensial, dll. Pemilihan pengklasifikasi seperti itu tetap tidak berubah terhadap biaya kemiringan dan kesalahan kelas yang berfungsi sebagai kondisi pengoperasian dalam pembelajaran yang sensitif terhadap biaya. Terakhir, semua kesimpulan yang diambil dari kurva ROC hanya relevan dalam sampel data pelatihan. Evaluasi kinerja terpisah dari model pembelajaran mesin diperlukan pada sampel data pengujian dan sampel data validasi. Di sini, sampel data adversarial dapat dianggap sebagai sampel data validasi.

Flach membahas metrik pembelajaran mesin yang mampu mengoptimalkan tradeoff akibat kelas yang miring dan kesalahan distribusi biaya klasifikasi yang diperoleh dari model pembelajaran mesin dalam pelatihan dan penerapan. Jenis plot kontur yang disebut plot isometrik ROC digunakan untuk menganalisis dan mengkarakterisasi perilaku berbagai metrik

pembelajaran mesin. Metrik pembelajaran mesin bersumber dari tabel kontingensi yang mentabulasi statistik kualitas model dalam kriteria pemisahan pohon keputusan, pola induksi aturan, klasifikasi, pengambilan informasi, dan penemuan subkelompok. Tingkat positif benar dan salah diasumsikan sebagai statistik yang memadai untuk mengkarakterisasi kinerja pengklasifikasi dalam konteks target apa pun. Metrik pembelajaran mesin yang diturunkan ditafsirkan untuk menambah makna statistik pada metrik konvensional. Misalnya, perhitungan akurasi diinterpretasikan dalam bentuk hasil yang diharapkan sebagai berikut:

i. Dengan mengabaikan biaya kesalahan klasifikasi, keakuratan memperkirakan kemungkinan bahwa contoh yang dipilih secara acak dapat diklasifikasikan dengan benar.
ii. Dengan biaya kesalahan klasifikasi, keakuratan memperkirakan kemungkinan bahwa contoh yang dipilih secara acak tidak menimbulkan biaya.
iii. Dengan biaya kesalahan klasifikasi dan keuntungan klasifikasi yang benar, keakuratan memperkirakan kemungkinan bahwa contoh yang dipilih secara acak menghasilkan keuntungan. Dengan demikian, kurva ROC dapat digunakan untuk mendeteksi sinyal diskriminatif dengan adanya gangguan adversarial dalam pembelajaran mendalam adversarial. Mereka dapat digunakan untuk menyempurnakan proposal pemodelan teoritis permainan yang mempertimbangkan biaya kesalahan klasifikasi. Mereka dapat digunakan untuk mengkarakterisasi perbedaan statistik antara sampel data pelatihan dan validasi di mana kumpulan data yang berlawanan bertindak sebagai sampel validasi yang mengubah distribusi kelas dalam sampel data pelatihan.

Ribeiro dkk. menjelaskan prediksi pemodelan dengan mempelajari model yang dapat ditafsirkan darinya. Ini memecahkan masalah optimasi submodular untuk memberikan pemodelan kepercayaan pengklasifikasi. Membangun kepercayaan pengguna terhadap model pembelajaran mesin adalah hal yang penting karena pengguna tidak akan menggunakan model kotak hitam atau keputusan preskriptif individu yang tidak mereka percayai.

Kepercayaan adalah perhatian penting dalam penggunaan langsung pengklasifikasi pembelajaran mesin sebagai alat serta penerapan model pembelajaran mesin dalam produk lain. Oleh karena itu, untuk mengevaluasi model pembelajaran mesin dalam kumpulan data dunia nyata, kami dapat merencanakan untuk memeriksa setiap prediksi dan penjelasannya yang bertindak sebagai metrik yang menarik dibandingkan rekayasa fitur yang menambah ukuran kinerja seperti akurasi, presisi, perolehan, dan skor-F.

Ribeiro dkk. mengidentifikasi karakteristik metode penjelasan yang diinginkan. Penjelasan tersebut mencakup penjelasan yang dapat diinterpretasikan yang memberikan pemahaman kualitatif antara variabel masukan dan respons, penjelasan ketelitian lokal yang sesuai dengan perilaku model di lingkungan sekitar kejadian yang diprediksi, penjelasan model-agnostik yang mampu memperlakukan model pembelajaran sebagai kotak hitam. , dan penjelasan perspektif global untuk memastikan kepercayaan terhadap model pembelajaran.

Goldstein dkk. membahas visualisasi algoritma kotak hitam dengan plot seperti plot ekspektasi kondisional individu (ICE) dan plot ketergantungan parsial (PDP). Teknik visualisasi

meliputi visualisasi batas keputusan dalam dimensi tinggi, visualisasi lapisan tersembunyi jaringan saraf untuk memahami ketergantungan antara masukan dan keluaran model dengan wawasan tentang ketidakpastian klasifikasi, representasi grafis dari kontribusi variabel terhadap kecocokan model dalam mesin vektor pendukung. , pendekatan teoretis permainan untuk menilai kontribusi berbagai fitur terhadap prediksi, dan estimasi kuasi-regresi fungsi kotak hitam.

Thiagarajan dkk. melakukan analisis sensitivitas jaringan saraf dalam. Ketidakpastian prediksi dianalisis dalam analisis sensitivitas dengan jaringan probabilistik. Tujuan pembelajaran adalah untuk menghasilkan model yang lebih kuat dan dapat digeneralisasikan tanpa mengorbankan interpretasi model. Regularisasi prediksi baru diperkenalkan untuk mendemonstrasikan jaringan saraf yang menggeneralisasi data yang tidak terlihat. Terakhir, ketidakpastian prediksi diuraikan dan dijelaskan dalam domain masukan untuk meningkatkan validasi dan interpretasi model pembelajaran mendalam. Gagasan tentang interpretabilitas dalam bab ini dibatasi pada pemahaman prediksi model dalam kaitannya dengan konstruksi sederhana yang dapat ditindaklanjuti pada fitur masukan.

Hasil eksperimen dibandingkan dengan pemodelan statistik konvensional yang mengadopsi alur inferensi Bayesian pada model terlatih untuk pembelajaran mendalam dengan estimasi titik. Pendekatan seperti itu mungkin tidak mampu menangani sampel uji yang tidak terdistribusi. Ketidakpastian prediksi diklasifikasikan menjadi ketidakpastian epistemik yang juga dikenal sebagai ketidakpastian model yang dapat dijelaskan dengan data pelatihan yang cukup dan ketidakpastian aleatorik yang bergantung pada noise atau keacakan dalam sampel masukan. Untuk mengatasi ketidakpastian aleatorik dengan asumsi distribusi input sebelumnya, penulis memasukkan estimasi mean dan varians untuk prediksi dalam analisis sensitivitas. Ketika ukuran sensitivitas tersebut digunakan sebagai pengatur tujuan pembelajaran, hal ini dikatakan akan menghasilkan kinerja generalisasi yang lebih baik.

Pada saat yang sama, fitur-fitur yang berkontribusi maksimal terhadap ketidakpastian model dilacak. Kemudian model yang memberikan ketidakpastian pada keluaran yang dapat diandalkan menunjukkan masalah pembelajaran baik dalam proses pelatihan atau data masukan. Keluaran pembelajaran mendalam diambil sebagai prediksi variabel respon regresi berkelanjutan. Fungsi kerugian berbasis kemungkinan bersyarat dipilih untuk melatih jaringan saraf menuju respons tersebut. Sensitivitas fitur dihitung dengan perluasan fungsi keputusan model Taylor orde pertama yang didekomposisi menjadi skor relevansi untuk setiap fitur masukan.

Sensitivitas fitur kemudian mengatur entropi kondisional yang berpusat pada parameter penting dalam fungsi kerugian dan proses pelatihan jaringan neural dalam. Dalam evaluasi eksperimental jaringan saraf dalam, pendekatan yang diusulkan menghasilkan peningkatan kinerja validasi dibandingkan dengan model dasar yang tidak memperhitungkan ketidakpastian. Pengaruh fitur masking insensitive dihitung dari statistik R-squared yang mengukur varians prediksi pada variabel dependen untuk regresi dari masing-masing variabel independen.

Zhang dkk. melakukan analisis sensitivitas jaringan saraf konvolusional satu lapisan (CNN) untuk klasifikasi kalimat. Tidak diketahui bagaimana CNN bergantung pada perubahan tak terduga pada representasi vektor kata masukan, ukuran wilayah filter, fungsi aktivasi, strategi pengumpulan, parameter regularisasi, jumlah peta fitur, hyperparameter, dan parameter bebas lainnya dalam arsitektur model. Ruang pencarian untuk semua kemungkinan arsitektur model sangatlah besar. SVM untuk klasifikasi kalimat digunakan sebagai model dasar untuk meningkatkan hasil CNN. Hasil eksperimen dapat digunakan untuk memandu teknik optimasi hyperparameter seperti pencarian grid, pencarian acak, dan optimasi Bayesian.

Sebagai bagian dari deteksi spam di dunia nyata, Pruthi et al. melakukan analisis sensitivitas pengenalan kata dengan jaringan saraf berulang (RNN) dengan adanya kesalahan ejaan yang berlawanan. Kesalahan ejaan yang dibuat secara adversarial terjadi dalam skenario serangan seperti menghapus, menambah, dan menukar karakter internal dalam masukan kata ke klasifikasi teks yang harus berurusan dengan pengeditan yang merugikan. Eksperimen menunjukkan bahwa musuh dapat menurunkan kinerja pengklasifikasi seperti yang dicapai dengan menebak secara acak. Untuk membatasi jumlah masukan yang berbeda pada pengklasifikasi, analisis sensitivitas mengurangi jumlah keluaran pengenalan kata yang berbeda yang dapat ditimbulkan oleh musuh.

Dengan demikian, tujuan pembelajaran adalah merancang sistem sensitivitas rendah dengan tingkat kesalahan yang rendah. Helton dkk. meninjau metode pengambilan sampel untuk analisis sensitivitas seperti pengambilan sampel acak, pengambilan sampel penting, dan pengambilan sampel hypercube Latin. Mereka membantu dalam konstruksi distribusi untuk mengkarakterisasi ketidakpastian stokastik dan subjektif dalam kumpulan data adversarial yang disebarkan melalui model pembelajaran mesin yang pada akhirnya memengaruhi prediksi model. Prosedur analisis sensitivitas yang ditinjau meliputi pemeriksaan scatterplot, analisis regresi, korelasi dan korelasi parsial, transformasi rank, serta identifikasi pola nonmonotonik dan nonrandom.

Xu dkk. memperoleh batasan generalisasi untuk algoritma pembelajaran mesin berdasarkan sifat ketahanannya. Di sini, kesalahan generalisasi dapat dipahami sebagai estimasi risiko algoritma pembelajaran. Hal ini diukur secara empiris dalam bentuk kesalahan kinerja pada kumpulan data pelatihan. Ukuran kompleksitas pada pembelajaran yang diawasi membatasi kesenjangan antara risiko yang diperkirakan dan risiko empiris dengan kompleksitas hipotesis yang ditetapkan untuk pembelajaran mesin. Diantaranya adalah dimensi Vapnik-Chervonenkis (VC), kompleksitas Kolmogorov, dan kompleksitas Rademacher. Xu dkk. mendefinisikan ketahanan algoritmik dengan mengacu pada tujuan optimasi minimum yang ditemukan dalam teori optimasi kuat.

Secara informal, algoritme pembelajaran dikatakan kuat jika mencapai performa yang “serupa” pada sampel pengujian dan sampel pelatihan yang “mendekati”. Banyak model pembelajaran mesin seperti regresi LASSO, mesin vektor pendukung, dan jaringan neural dalam dapat diformulasi ulang agar memiliki tujuan pembelajaran dalam kerangka pengoptimalan yang kuat untuk menargetkan minimalisasi kesalahan kinerja empiris dalam

gangguan masukan terburuk yang mungkin terjadi dalam beberapa kumpulan ketidakpastian yang ditentukan dengan benar untuk optimasi. Di sini, kemampuan generalisasi algoritma pembelajaran dapat diselidiki dalam kaitannya dengan nilai fungsi kerugian yang diharapkan dari hipotesis yang dipelajari pada sampel yang secara statistik menyimpang dari sampel pelatihan.

Dalam pengaturan pelatihan adversarial, kerugian yang diharapkan tersebut disesuaikan untuk meminimalkan manipulasi fitur dalam pembelajaran adversarial dan biaya kesalahan klasifikasi dalam pemodelan teoritis permainan. Dalam analisis model pembelajaran mesin lainnya, perkiraan kerugian dapat disesuaikan untuk pembelajaran metrik, pembelajaran transfer, pembelajaran penguatan, dan pembelajaran dengan outlier. Batasan generalisasi pada kerugian yang diharapkan yang berasal dari kerangka ketahanan algoritmik yang diusulkan dapat menangani pengaturan pembelajaran transfer karena kumpulan data yang tidak cocok dalam adaptasi domain. Hal ini dapat diperluas ke penyelidikan mengenai ketahanan algoritma pembelajaran tanpa pengawasan dan semi-pengawasan.

Untuk memperbaiki kesalahan dalam pengaturan klasifikasi, Asif et al. membuat pengklasifikasi yang sensitif terhadap biaya yang dapat dikenakan sanksi pada label kelas prediksi dan aktual yang bergantung pada aplikasi. Pendekatan teoritis permainan min-maks yang kuat menghasilkan pengklasifikasi yang meminimalkan biaya kesalahan dalam klasifikasi sebagai masalah optimasi cembung. Pengoptimalan seperti itu dapat dilakukan dibandingkan dengan pendekatan minimalisasi risiko empiris NP-hard yang mengatasi kesalahan biaya dalam klasifikasi sensitif biaya dengan fungsi kerugian non-cembung. Ini adalah pendekatan untuk meminimalkan biaya yang diharapkan dalam pembelajaran mesin yang tangguh. Ini secara langsung meminimalkan kerugian yang sensitif terhadap biaya pada perkiraan data pelatihan.

Permainan zero-sum yang diusulkan diselesaikan dengan menggunakan program linier. Evaluasi kinerja pembelajaran mesin dilakukan pada matriks biaya kebingungan. Sensitivitas biaya dapat digunakan untuk mempertimbangkan kembali data pelatihan yang tersedia, memasukkan biaya kebingungan ke dalam formulasi pengklasifikasi, dan meningkatkan kumpulan pengklasifikasi yang lemah untuk menghasilkan pembelajar yang sensitif terhadap biaya. Lebih lanjut, fungsi kerugian untuk proses pelatihan dapat secara langsung memasukkan sensitivitas biaya ke dalam generalisasi multikelas dari pengklasifikasi biner. Perspektif pembelajaran yang berlawanan mengenai sensitivitas biaya membawa dimensi tambahan pada pemodelan klasifikasi, estimasi statistik, dan pengambilan keputusan dalam kondisi ketidakpastian.

Di sini, metode pembelajaran adversarial yang relevan mencakup model pengambilan keputusan maximin sebagai permainan adversarial berurutan, optimasi mini-max dari penyesalan keputusan, estimasi statistik dalam ketidakpastian yang meminimalkan risiko kasus terburuk, dan model entropi maksimum yang menggunakan kerugian logaritmik. distribusi keluarga eksponensial. Distribusi probabilitas diperkirakan sebagai solusi dari permainan min-maks tersebut. Pembelajaran yang sensitif terhadap biaya yang

menggabungkan pembelajaran adversarial menjadi lebih kuat tidak hanya terhadap pergeseran distribusi dalam kumpulan data tetapi juga terhadap ketidakpastian karena distribusi bersyarat pada label dalam fungsi kerugian.

Tanpa mengasumsikan persamaan bentuk tertutup dalam bentuk parametrik untuk data tertentu, pendekatan ini memungkinkan kita untuk menggabungkan properti data pelatihan dan distribusi data bersyarat sebagai batasan klasifikasi karena distribusi label bersyarat musuh. Melihat tugas klasifikasi yang sensitif terhadap biaya sebagai permainan dua pemain antara estimator dan musuh membatasi musuh untuk memilih distribusi manipulasi data yang cocok dengan statistik vektor momen dari distribusi input yang mendasarinya. Kompleksitas komputasi estimator secara implisit bertambah seiring dengan besarnya dimensi batasan tersebut. Pemilihan dan regularisasi fitur yang cermat dapat menghindari masalah seperti itu.

Lundberg dkk. menyajikan kerangka kerja yang disebut SHAP (SHapley Additive exPlanations) untuk menafsirkan prediksi dalam pembelajaran mendalam. Penjelasan tentang prediksi suatu model dianggap sebagai model itu sendiri. Ini disebut model penjelasan dan mendefinisikan kelas metode atribusi fitur aditif. Pemodelan teoretis permainan kemudian menjamin solusi unik untuk seluruh kelas metode atribusi fitur aditif. Model penjelasan tersebut menggunakan masukan yang disederhanakan yang dipetakan ke masukan asli melalui fungsi pemetaan. Mereka diselesaikan sebagai model regresi linier yang diberi sanksi.

Nilai regresi Shapley dari teori permainan kooperatif digunakan untuk menemukan pentingnya fitur untuk model linier dengan adanya multikolinearitas. Samek dkk. memperkenalkan perlunya kecerdasan buatan yang dapat dijelaskan dalam domain AI seperti klasifikasi gambar, analisis sentimen, pemahaman ucapan, dan permainan strategis. Perkembangan terkini untuk memvisualisasikan, menjelaskan, dan menafsirkan model pembelajaran mendalam kemudian disurvei. Dua metode analisis sensitivitas disajikan untuk menjelaskan prediksi dalam pengklasifikasi pembelajaran mendalam kotak hitam. Salah satu metode menghitung sensitivitas prediksi sehubungan dengan perubahan masukan. Metode lain menguraikan keputusan berdasarkan variabel masukan. Das dkk. mengulas lanskap kecerdasan buatan yang dapat dijelaskan (XAI).

Taksonomi teknik XAI disediakan. Penggunaannya untuk membangun model pembelajaran mendalam yang dapat dipercaya, dapat ditafsirkan, dan cukup jelas telah disurvei. Selain penciptaan contoh adversarial dalam keputusan pengklasifikasi yang menyesatkan, XAI harus memiliki realisasi rekayasa fitur yang berpusat pada alasan etika, peradilan, dan keamanan. “Interpretabilitas” didefinisikan sebagai kualitas atau fitur yang diinginkan dari suatu algoritme yang menyediakan data ekspresif yang cukup untuk memahami cara kerja algoritme. “Interpretasi” didefinisikan sebagai representasi sederhana dari domain kompleks, seperti keluaran yang dihasilkan oleh model pembelajaran mesin, menjadi konsep bermakna yang dapat dipahami dan masuk akal oleh manusia.

Sebuah “penjelasan” didefinisikan sebagai informasi meta tambahan, yang dihasilkan oleh algoritme eksternal atau oleh model pembelajaran mesin itu sendiri, untuk

mendeskripsikan pentingnya fitur atau relevansi instance masukan terhadap klasifikasi keluaran tertentu. Setiap penjelasan harus konsisten pada titik data yang serupa dan menghasilkan penjelasan yang stabil atau serupa pada titik data yang sama dari waktu ke waktu. Jadi Das dkk. pelajari XAI dalam sistem produksi pembelajaran mesin untuk mengetahui kepercayaan, transparansi, bias, dan keadilan.

Model pembelajaran mendalam dianggap “transparan” jika cukup ekspresif sehingga dapat dipahami manusia. Di sini, transparansi dapat menjadi bagian dari algoritma itu sendiri atau menggunakan cara eksternal seperti dekomposisi model atau simulasi. “Kepercayaan” model pembelajaran mendalam adalah ukuran keyakinan, sebagai manusia, sebagai pengguna akhir, terhadap tujuan kerja model tertentu dalam lingkungan dunia nyata yang dinamis. “Keadilan” dalam pembelajaran mendalam adalah kualitas model yang dipelajari dalam memberikan keputusan yang tidak memihak dan adil tanpa memihak populasi mana pun dalam distribusi data masukan. Keadilan memitigasi bias yang ditimbulkan pada keputusan AI baik dari kumpulan data masukan atau arsitektur jaringan saraf yang buruk.

Pembelajaran adversarial teori permainan yang diusulkan dalam buku ini dapat digunakan untuk mengembangkan algoritma komputasi untuk tujuan optimasi dan inferensi statistik dalam kapasitas algoritma pembelajaran adversarial untuk pengacakan, diskriminasi, keandalan, dan kemampuan belajar. Mempelajari kompleksitas komputasi pemodelan teoritis permainan dalam pembelajaran adversarial mengakomodasi perluasan penelitian pada ketahanan, keadilan, kemampuan menjelaskan, dan transparansi model pembelajaran mesin. Kita dapat mensimulasikan pengkodean variasional dari batas-batas keputusan yang dapat dipelajari yang dihasilkan dari pembelajaran mendalam adversarial teoritis permainan sebagai masalah penyimpanan-pengambilan dalam penambangan data pada manipulasi adversarial.

Keandalan pembelajaran mesin dalam penerapan dapat disimulasikan dengan optimasi komputasi dan masalah inferensi statistik dalam analisis lanjutan dari musuh teoritis permainan dalam pembelajaran mendalam. Selain kendala operasi dalam kebijakan keamanan, kendala jarak dan anggaran dalam fungsi biaya adversarial menjadi perhatian penelitian kami. Di sini, teori permainan yang digerakkan oleh batasan dan komputasi evolusioner diperlukan untuk menyelesaikan masalah optimasi multi-tujuan, terbatas, berskala besar, dan tidak pasti dalam skenario serangan kotak hitam.

Kesulitan komputasi lebih lanjut untuk mengukur utilitas dan hilangnya informasi terkait dapat diatasi dalam formulasi teori permainan. Kemudian pembelajaran adversarial berbasis teori keputusan memberikan analisis data adaptif. Setelah melatih algoritma pembelajaran pada kumpulan data yang signifikan secara statistik yang dihasilkan oleh lawan teori permainan, pembelajaran mendalam dapat menskalakan dan memvalidasi pemodelan pembelajaran mesin ke dalam pengaturan data besar dengan model komputasi dari pengambilan keputusan yang dilakukan secara human-in-the-loop. Model yang kuat untuk pengambilan keputusan berdasarkan data dalam pembelajaran mesin adversarial teoretis game mengasumsikan tersedianya informasi yang tidak sempurna untuk mempelajari parameter pemodelan.

Mereka mengoptimalkan distribusi probabilitas pada data yang tidak pasti untuk menghindari kesalahan estimasi. Dalam hal pengoptimalan yang kuat, variabel acak yang mendasari fitur pembelajaran mesin dimodelkan sebagai parameter ketidakpastian yang termasuk dalam kumpulan ketidakpastian cembung. Pengambil keputusan berdasarkan data kemudian dilindungi dengan sistem pembelajaran mesin yang dibangun untuk menghadapi skenario terburuk dalam rangkaian tersebut. Pendekatan untuk membangun batasan ketahanan dalam pembelajaran mendalam adversarial harus bersaing dengan contoh-contoh adversarial yang dirancang untuk menyesatkan klasifikasi gambar agar melakukan tindakan yang tidak diinginkan. Oleh karena itu, kekuatan pengambilan keputusan berbasis data dapat diteliti dalam pembelajaran mesin adversarial. Kami kemudian dapat merancang fungsi kerugian dalam pembelajaran mendalam dengan tujuan teoritis permainan.

BAB 4
TEORITIS PERMAINAN DALAM MANIPULASI PEMBELAJARAN MESIN MENDALAM

Bab ini merangkum strategi teoritis permainan untuk menghasilkan manipulasi permusuhan. Tujuan pembelajaran adversarial untuk musuh kita diasumsikan untuk memasukkan perubahan kecil ke dalam distribusi data, yang didefinisikan pada label kelas positif dan negatif, hingga pembelajaran mendalam kemudian menyebabkan kesalahan klasifikasi pada distribusi data. Dengan demikian, tujuan teoritis dari proses pembelajaran mendalam permusuhan kita menjadi salah satu penentuan apakah manipulasi data masukan telah mencapai batas keputusan pembelajar, yaitu di mana terlalu banyak label positif menjadi label negatif.

Data permusuhan dihasilkan dengan menyelesaikan kebijakan serangan yang optimal dalam permainan Stackelberg di mana musuh menargetkan kinerja kesalahan klasifikasi pembelajaran mendalam. Formulasi teoritis permainan sekuensial dapat memodelkan interaksi antara musuh yang cerdas dan model pembelajaran mendalam untuk menghasilkan manipulasi permusuhan dengan menyelesaikan permainan Stackelberg non-kooperatif berurutan dua pemain di mana fungsi pembayaran setiap pemain meningkat dengan interaksi ke optimal lokal. Dengan rumusan teori permainan stokastik, kita kemudian dapat memperluas permainan Stackelberg dua pemain menjadi permainan Stackelberg multipemain dengan fungsi pembayaran stokastik untuk lawan. Kedua versi permainan tersebut diselesaikan melalui ekuilibrium Nash, yang mengacu pada sepasang strategi di mana tidak ada insentif bagi pembelajar atau lawan untuk menyimpang dari strategi optimal mereka.

Kami kemudian dapat mengeksplorasi musuh yang mengoptimalkan fungsi pembayaran variasional melalui strategi pengacakan data pada pembelajaran mendalam yang dirancang untuk tugas klasifikasi multi-label. Demikian pula, hasil dari penyelidikan ini adalah desain algoritme yang memecahkan permainan Stackelberg sekuensial dua pemain dengan jumlah variabel dengan keseimbangan Nash yang baru. Musuh memanipulasi parameter variasi dalam data masukan untuk menyesatkan proses pembelajaran pembelajaran mendalam, sehingga salah mengklasifikasikan label kelas asli sebagai label kelas yang ditargetkan. Manipulasi adversarial variasional yang ideal adalah perubahan minimum yang diperlukan pada fungsi biaya adversarial dari data yang dikodekan yang akan mengakibatkan pembelajaran mendalam salah memberi label pada data yang didekodekan.

Manipulasi optimal disebabkan oleh stokastik optima pada strategi respon terbaik non-cembung. Data permusuhan yang dihasilkan oleh varian permainan Stackelberg ini mensimulasikan interaksi berkelanjutan dengan proses pembelajaran pengklasifikasi dibandingkan dengan interaksi satu kali. Proses pembelajaran CNN dapat dimanipulasi oleh musuh pada tingkat data masukan serta tingkat data yang dihasilkan. Kami kemudian dapat

melatih kembali model pembelajaran mendalam yang asli pada data yang dimanipulasi untuk menghasilkan model pembelajaran mendalam adversarial yang aman dan tahan terhadap kerentanan kinerja selanjutnya dari musuh teoretis game. Hipotesis alternatif untuk penambangan data adversarial dalam permainan strategi pembelajaran mendalam adversarial teoretis disediakan dalam aplikasi keamanan siber dengan pembelajaran mesin yang dirancang untuk persyaratan keamanan. Konsep solusi teoritis permainan mengarah pada jaringan saraf dalam yang kuat terhadap manipulasi data selanjutnya oleh musuh teoritis permainan. Hasil yang menjanjikan ini menunjukkan bahwa algoritma pembelajaran berdasarkan pemodelan teori permainan dan optimasi matematika adalah pendekatan yang jauh lebih baik untuk membangun model pembelajaran mendalam yang lebih aman.

Permainan stokastik yang mendefinisikan ruang strategi untuk manipulasi permusuhan telah digunakan untuk menghasilkan contoh permusuhan. Ruang strategi seperti itu didefinisikan dalam bentuk tindakan dua atau lebih pihak yang berlawanan dan fungsi imbalan yang sesuai. Masing-masing musuh tersebut dapat melibatkan satu atau lebih pembelajar dalam sebuah permainan dan sebaliknya. Dari sudut pandang pelajar, menyesuaikan parameter model teoretis permainan secara komputasi lebih murah dibandingkan membuat model baru yang tahan terhadap manipulasi permusuhan. Dari sudut pandang musuh, skenario serangan dapat dicirikan oleh parameter optimasi stokastik yang diperkirakan dalam interaksi teoretis permainan dengan pelajar.

4.1 MODEL PEMBELAJARAN TEORI PERMAINAN

Ide permainan sekuensial dua pemain (atau permainan Stackelberg) dan permainan kooperatif multipemain telah digunakan sebagai kerangka teori permainan yang melatih algoritma pembelajaran permusuhan. Mencari keseimbangan dalam permainan seperti itu setara dengan memecahkan masalah optimasi berdimensi tinggi. Performa model akhirnya kemudian diperkirakan dengan metode optimasi stokastik berdasarkan algoritma pencarian heuristik yang efisien secara komputasi. Selama fungsi tujuan dibatasi, metode optimasi global seperti algoritma genetika, simulasi anil, dan pendakian bukit stokastik dapat diterapkan untuk mencari kriteria konvergensi yang mengarah pada keseimbangan sempurna subgame.

Globerson dkk. membahas algoritma klasifikasi dengan rumusan teori permainan. Algoritme yang diusulkan kuat untuk menghapus fitur berdasarkan fungsi tujuan min-maks yang dioptimalkan dengan pemrograman kuadratik. Dalam Liu dkk., interaksi antara musuh dan penambang data dimodelkan sebagai permainan zero-sum Stackelberg berurutan dua pemain di mana hasil untuk setiap pemain dirancang sebagai fungsi kerugian yang diatur. Musuh secara berulang menyerang penambang data menggunakan strategi terbaik untuk mengubah data pelatihan asli.

Penambang data bereaksi secara independen dengan membangun kembali pengklasifikasi berdasarkan pengamatan penambang data terhadap modifikasi musuh pada data pelatihan. Permainan seperti itu diulangi sampai hadiah lawan tidak bertambah atau jumlah iterasi maksimum tercapai. Liu dkk. mengusulkan perpanjangan ke Liu dkk. di mana

permainan satu langkah digunakan untuk mengurangi waktu komputasi algoritma minimax. Metode satu langkah konvergen ke ekuilibrium Nash dengan memanfaatkan dekomposisi nilai tunggal (SVD). Liu dkk. merumuskan masalah optimasi bilevel dari permainan jumlah bukan nol pada transformasi data permusuhan. Game ini bereksperimen dengan pengatur renggang untuk merancang tujuan klasifikasi yang kuat.

Sebuah permainan berakhir dalam keseimbangan dengan imbalan kepada setiap pemain berdasarkan tujuan dan tindakan mereka. Pembelajar tidak mempunyai insentif untuk memainkan permainan yang menghasilkan terlalu banyak positif palsu dan terlalu sedikit peningkatan dalam positif sebenarnya. Musuh tidak memiliki insentif untuk memainkan permainan yang meningkatkan kegunaan negatif palsu yang tidak terdeteksi oleh algoritma pembelajaran. Pada kondisi keseimbangan, musuh mampu menemukan data pengujian yang berbeda secara signifikan dari data pelatihan, sedangkan pembelajar mampu memperbarui modelnya untuk menghadapi ancaman baru dari data musuh.

Semua pemain diasumsikan bertindak sesuai kepentingan rasionalnya untuk memaksimalkan keuntungan. Asumsi ini, pada setiap tahapan permainan, menghilangkan keseimbangan Nash dengan ancaman yang tidak dapat dipercaya terhadap pelajar dan menciptakan keseimbangan yang disebut keseimbangan sempurna subgame. Di sini, keseimbangan sempurna mengasumsikan bahwa masing-masing pemain mengetahui fungsi utilitas pihak lain. Fungsi utilitas pemain berbeda-beda menurut domain aplikasi.

Dasar-dasar Teori Permainan

Teori permainan menyediakan alat matematika untuk memodelkan perilaku pembela dan perilaku musuh dalam pembelajaran mesin dalam hal strategi pertahanan dan serangan. Pembelajaran permusuhan teoritis permainan memperhitungkan trade-off yang dilakukan oleh penyerang antara biaya adaptasi terhadap pengklasifikasi dan manfaat dari serangan tersebut. Di sisi lain, pengorbanan yang dilakukan oleh pembela HAM menyeimbangkan antara manfaat deteksi serangan yang benar dan kerugian jika terjadi alarm palsu. Optima dalam pembelajaran adversarial mampu menentukan strategi apa yang cocok diperlukan untuk mengurangi kerugian bek dari serangan adversarial.

Interaksi strategis antara fungsi pembayaran untuk kedua pemain mencerminkan peringkat relatif skenario aplikasi masing-masing pemain dalam hal hasil akhir yang diharapkan dalam pembelajaran mesin. Permainan Stackelberg biasanya digunakan untuk memodelkan interaksi strategis dengan asumsi agen rasional di pasar di mana terdapat persaingan hierarkis. Ruang pencarian strategi untuk setiap pemain dalam permainan biasanya diasumsikan berbatas dan cembung, dan fungsi pembayaran yang sesuai diasumsikan dapat dibedakan. Solusi ekuilibrium untuk semua fungsi pembayaran dalam permainan ditentukan oleh solusi fungsi tujuan optimasi. Teori permainan dapat diterapkan dalam bidang ekonomi, teori politik, ilmu evolusi, dan strategi militer.

Webb dkk. menyajikan pengantar teori permainan. Ini mencakup model keputusan dan proses pengambilan keputusan untuk menentukan agen rasional yang berpartisipasi dalam permainan statis, permainan dinamis, dan permainan evolusi. Berbagai gagasan tentang keseimbangan teoretis permainan mengarah pada gagasan deskriptif atau preskriptif

tentang analisis data. Dalam merumuskan permainan, kita harus mendefinisikan pemain, tindakan dan informasi yang tersedia bagi para pemain, informasi waktu tentang interaksi antar pemain baik secara simultan atau berurutan, urutan permainan dan pengulangan interaksi, imbalan kepada berbagai pemain sebagai hasilnya. interaksi, dan estimasi biaya-manfaat dari setiap rangkaian pilihan potensial untuk semua pemain. Osborne dkk. merancang buku teks untuk kursus pascasarjana dalam teori permainan. Ini mencakup permainan strategis, permainan ekstensif, dan permainan koalisi.

Leyton-Brown dkk. mendefinisikan teori permainan sebagai studi matematis tentang interaksi antara agen-agen yang independen dan mementingkan diri sendiri. Kelas utama permainan, representasinya, dan konsep utama yang digunakan untuk menganalisisnya dirangkum. Teori utilitas dikembangkan untuk memodelkan minat dan preferensi agen pada serangkaian alternatif yang tersedia dalam permainan seperti permainan bentuk normal, permainan bentuk ekstensif, permainan informasi tidak sempurna, permainan berulang, permainan stokastik, permainan Bayesian, dan permainan koalisi. Agen yang dihadapkan pada ketidakpastian dalam lingkungan pembelajaran kemudian menentukan nilai yang diharapkan dari fungsi utilitas sehubungan dengan distribusi probabilitas yang sesuai di seluruh negara bagian.

Secara sederhana, utilitas dapat diartikan sebagai besarnya kebahagiaan yang diperoleh seorang agen (pemain) dari suatu hasil atau imbalan tertentu. Hasil teoritis permainan yang menarik dalam sistem pembelajaran mesin dapat dikategorikan dalam subkumpulan hasil yang mungkin sebagai konsep solusi seperti yang ditemukan karena optimalitas Pareto dan keseimbangan Nash. Jadi permainan strategis dalam pembelajaran mesin harus memodelkan komponen pembelajaran permusuhan seperti sekumpulan pemain, serangkaian strategi untuk setiap pemain, dan fungsi pembayaran yang menunjukkan hasil yang diinginkan dalam permainan bagi seorang pemain.

Konteks teoretis permainan (karena keterampilan atau strategi) pengambilan keputusan untuk setiap pemain berguna untuk menganalisis pengambilan keputusan berdasarkan data yang dibuat oleh sistem pembelajaran mesin yang berisiko. Ini telah diterapkan pada ilmu-ilmu sosial untuk menciptakan teori pilihan rasional sebagai adaptasi dari filosofi individualisme metodologis untuk memaksimalkan utilitas/mata uang/nilai tindakan individu di antara perilaku kolektif. Suatu pilihan dianggap “rasional” dalam ilmu ekonomi jika hal tersebut mengarah pada peringkat preferensi atas serangkaian item yang mencirikan alternatif bagi pengambil keputusan dimana semua perbandingannya konsisten. Gagasan konsistensi yang lebih maju harus memperhitungkan ketidakpastian dalam lingkungan belajar dan pengambilan keputusan dari waktu ke waktu ketika seorang pemain tidak memiliki informasi yang tepat dan kemampuan kognitif tentang hasil dari pilihan dan perbandingan di antara keduanya. Jadi proses pengambilan keputusan dalam teori pilihan rasional harus divalidasi secara empiris untuk ide-ide seperti “alasan”, “preferensi”, “rasionalitas”, dan “kemampuan belajar” dengan sifat matematika formal yang berguna dalam sistem pembelajaran mesin.

Bergantung pada interaksi pemain dalam tujuan teoretis permainan, solusi kesetimbangan disebut kesetimbangan Stackelberg atau kesetimbangan Nash. Dalam penelitian kami, kami menerapkan pemodelan teoretis permainan untuk masalah pembelajaran mesin yang diawasi yang menyimpulkan batasan keputusan dan distribusi data yang sesuai dalam sampel data pelatihan dan sampel data validasi. Kami mewakili biaya partisipasi dalam game dalam hal kinerja kesalahan klasifikasi dan biaya pelatihan ulang dalam pembelajaran mendalam.

Algoritma pencarian evolusioner dan optimasi digunakan untuk menyelesaikan permainan untuk menemukan manipulasi permusuhan terhadap data pelatihan. Bergantung pada pentingnya biaya yang dikeluarkan untuk menghasilkan serangan dan untuk melatih kembali pengklasifikasi, kami menemukan solusi keseimbangan yang berbeda untuk permainan tersebut. Selanjutnya, kami mengasumsikan skenario serangan kotak hitam di mana musuh tidak dapat mengamati strategi pengklasifikasi sebelum memilih strateginya. Kami kemudian bereksperimen dengan varian dalam skenario serangan di mana kerugian utilitas pemain bertahan dalam permainan lebih rendah dibandingkan utilitas musuh dalam permainan non-zero sum.

Penelitian kami meluas ke model permainan Bayesian di mana pemain memiliki informasi yang tidak lengkap tentang pemain lain. Hal ini lebih mungkin terjadi karena pembela HAM mungkin tidak mengetahui biaya pasti untuk menghasilkan data permusuhan dan penyerang mungkin tidak mengetahui biaya klasifikasi yang tepat untuk pembela HAM. Mereka hanya mempunyai keyakinan mengenai biaya-biaya tersebut. Pendekatan pemodelan ini mengubah permainan informasi yang tidak lengkap menjadi permainan informasi yang tidak sempurna. Dengan demikian, teknik pembelajaran permusuhan yang mengandalkan kerangka kerja berbasis teori permainan bisa menjadi relevan karena memodelkan perilaku pembelajar dan musuh berdasarkan manfaat dan biaya yang dikeluarkan untuk melatih ulang model dan menghasilkan penyerang.

Melatih pengklasifikasi dengan contoh permusuhan dalam data permusuhan sintetik mirip dengan regularisasi pengklasifikasi. Dalam konteks ini, teori permainan memberikan alat yang berguna untuk memodelkan perilaku musuh dan pembelajar karena teori ini mencakup, di satu sisi, manfaat bagi musuh untuk menyerang dan biaya untuk menghasilkan data permusuhan, dan di sisi lain, biaya pelajar untuk memperbarui model. Oleh karena itu, pendekatan berbasis teori permainan menyoroti trade-off yang dilakukan pihak lawan dan pembelajar dan dapat digunakan untuk menilai risiko penerapan teknologi keamanan siber tertentu untuk pengambilan keputusan berbasis data. Dari sudut pandang komputasi, prosedur pengambilan keputusan dapat dikodekan ke dalam algoritma dan heuristik yang mensimulasikan rasionalitas secara efektif.

Salah satu cara penting untuk mempelajari rasionalitas adalah dengan mengusulkan agen berdasarkan asumsi yang diadopsi oleh algoritma dan heuristik yang berbeda. Kita kemudian dapat mempelajari keseimbangan pasar interaksi antara agen-agen seperti kuantifikasi dampak algoritma dan heuristik dalam model analisis data. Teori permainan dapat digunakan untuk menurunkan keseimbangan pasar secara matematis. Studi berbasis

agen tersebut juga melibatkan kecerdasan komputasi. Karena ledakan kombinatorial dari algoritma optimal dan heuristik biasanya sulit dilakukan secara komputasi, kemampuan kita untuk secara efektif menggunakan kekuatan komputasi yang tersedia untuk menemukan solusi yang baik ditentukan oleh algoritma kecerdasan komputasi yang kita terapkan.

Tingkat optimalitas yang dapat kita capai secara wajar dalam paradigma berbasis agen kemudian menentukan rasionalitas efektif kita dalam masalah pembelajaran mesin. Prosedur pengambilan keputusan dalam kecerdasan komputasi dapat dikembangkan dengan algoritma pembelajaran evolusioner. Mereka mampu memisahkan pengetahuan spesifik domain dari mekanisme penalaran. Dengan demikian, rasionalitas di dunia nyata dapat dipelajari dalam konteks masalah pengambilan keputusan dalam kecerdasan komputasi. Dengan visualisasi transformasi data dalam kecerdasan komputasi termasuk analisis data gangguan, fungsi keanggotaan fuzzy dapat dirancang untuk mengurangi efek outlier dan gangguan pada batasan keputusan klasifikasi dengan memberikan penalti kesalahan pelatihan secara berbeda.

Pemecah masalah umum dalam kecerdasan komputasi adalah bidang kecerdasan buatan dalam perencanaan. Mereka melibatkan representasi pengetahuan tentang keyakinan, tindakan, dan dampaknya, penalaran kausal tentang tindakan dan konsekuensinya, dan alokasi sumber daya tentang penentuan waktu kinerja di seluruh tindakan. Masalah keputusan pilihan terbatas adalah topik kepuasan kendala yang merupakan titik temu antara kecerdasan buatan, pemrograman logika, dan riset operasi. Tampaknya dalam aplikasi seperti penjadwalan industri dan perencanaan produksi. Algoritma pencarian dan optimasi dapat dirancang untuk menggunakan kepuasan kendala untuk menemukan solusi yang efisien.

Algoritme kecerdasan komputasi yang disebut prosedur pencarian lokal dapat meniru pemikiran strategis manusia. Mereka dapat mengotomatiskan perbaikan berulang terhadap situasi saat ini, mencari kemungkinan perubahan secara eksperimental, dan mengubah strategi sebagai respons terhadap perubahan aktual atau yang diantisipasi. Prosedur dinamis untuk penelusuran lokal dapat memodelkan rasionalitas manusia dan pembelajaran penguatan dengan algoritme evolusioner yang mengembangkan solusi alih-alih merancangnya. Prosedur tersebut dapat berupa prosedur komputasi evolusioner atau prosedur yang dihasilkan oleh komputasi evolusioner. Dengan menentukan prosedur stasioner dan evolusioner di agen yang berbeda, kami dapat mengidentifikasi strategi perdagangan dan perilaku pasar. Representasi, penjelasan, penalaran, dan pembelajaran dengan agen intelijen komputasi harus memastikan kesesuaian statistik untuk algoritma kecerdasan komputasi.

Camerer dkk. menggambarkan teori permainan sebagai sistem matematika untuk menganalisis dan memprediksi situasi strategis. Keseimbangan teoretis permainan didasarkan pada pemikiran strategis yang saling konsisten dan strategi respons terbaik dalam keseimbangan tersebut. Respons terbaik ditentukan oleh keyakinan pada optimalisasi permukaan serangan musuh. Sistem pembelajaran mesin diasumsikan bertindak rasional dalam situasi tertentu. Kehadirannya sebagai pemain yang rasional kemudian mengubah

permainan teoritis optimal yang dicapai oleh para pemain dan populasi musuh yang heterogen. Model pembelajaran penguatan diusulkan untuk menganalisis secara empiris kekuatan prediksi permainan berulang. Contoh adversarial diartikan sebagai contoh tandingan dan penyimpangan dari kriteria keberhasilan. Mereka mampu mengubah parameterisasi rasionalitas teoretis permainan secara dinamis.

Penyempurnaan dan pemilihan strategi teoritis permainan seorang pemain tunduk pada keyakinan tentang strategi pemain yang tersisa dalam permainan kemudian ditemukan solusi respons stokastik yang lebih baik dalam jalur menuju keseimbangan statistik suatu permainan. Rasionalitas yang dapat diamati dalam pengambilan keputusan teoritis permainan bagi pembelajar dan lawan kemudian mengarah pada konsep "rasionalitas terbatas" yang memiliki informasi tentang keputusan, imbalan, kemampuan komputasi, dan kecerdasan dalam rumusan teori pembelajaran permainan. tujuan yang dirancang untuk menghasilkan perilaku fungsional untuk setiap pemain. Kelas konsep teori permainan yang dihasilkan terdiri dari sistem pembelajaran mesin dengan kemampuan untuk belajar dan berkembang seiring waktu.

Analisis "kasus yang masuk akal" terhadap musuh dan batasan optimasinya mengarah pada evaluasi keamanan kuantitatif dari metode pembelajaran mesin yang dinyatakan dalam bentuk komputasi serangan optimal dan penurunan batas atas risiko permusuhan untuk algoritma pembelajaran statistik. . Di sini, kerangka penalaran yang dibangun berdasarkan representasi pengetahuan dari contoh-contoh permusuhan mencakup penalaran berbasis kasus, penalaran berbasis aturan, dan penalaran berbasis data. Mereka tidak hanya dapat memberikan strategi optimasi maksimal untuk algoritma pembelajaran tetapi juga menyediakan kriteria teoritis permainan untuk desain sistem dengan pemrograman matematis pada fungsi kebugaran yang terputus-putus dan berisik yang memiliki berbagai jenis desain, sistem, dan kendala operasional.

Selain kendala operasi dalam kebijakan keamanan, kendala jarak dan anggaran dalam fungsi biaya adversarial menjadi perhatian penelitian kami. Kepercayaan pembelajaran mesin dalam penerapan dapat disimulasikan dengan optimasi komputasi dan masalah inferensi statistik dalam analisis tingkat lanjut dengan pembelajaran mendalam adversarial teoretis permainan. Ini juga dapat menggabungkan kontrol sistem dinamis dalam optimalisasi black-box pembelajaran mendalam. Di sini, teori permainan yang digerakkan oleh batasan dan komputasi evolusioner diperlukan untuk menyelesaikan masalah optimasi multi-tujuan, terbatas, berskala besar, dan tidak pasti. Kendala spesifik aplikasi ditentukan oleh pengambilan keputusan dalam data mining. Dengan memodelkan kebocoran informasi sebagai fungsi kerugian dalam pembelajaran mendalam, solusi pengoptimalan kami dalam keseimbangan teoretis game mampu merumuskan kebocoran informasi dari informasi pribadi yang tersedia di platform AI sebagai pengaturan permusuhan. Kami juga melakukan studi terhadap fungsi biaya adversarial yang ada sehubungan dengan batasan ketahanan dan anggaran privasi dalam model pembelajaran representasi mendalam untuk pembelajaran adversarial.

Kami ingin mencapai konvergensi terhadap perkiraan distribusi target dengan pembelajaran generatif yang mendalam. Sejauh mana kebisingan permusuhan dapat bermanfaat bagi proses pelatihan serta kualitas keseluruhan distribusi yang dihasilkan oleh pembelajaran permusuhan teoritis permainan bergantung pada sifat spesifik dari distribusi target yang dihasilkan. Di sini, kita dapat membingkai ekstensi penambangan data dari pembelajaran mendalam yang bermusuhan ke dalam penambangan data web, analisis deret waktu, sistem cyber-fisik, manipulasi sistem otonom, pengenalan pola multimedia, dan analisis keamanan jaringan.

Sebagai model pemikiran yang terbatas di bidang ekonomi, kerangka pemikiran tersebut dapat diterapkan dalam menjelaskan gelembung harga, spekulasi dan pertaruhan, pengabaian persaingan dalam strategi bisnis, kesederhanaan kontrak insentif, dan persistensi guncangan nominal dalam perekonomian makro. Algoritme pembelajaran yang dihasilkan dapat diterapkan dalam menjelaskan evolusi penetapan harga, kontrak berulang, organisasi industri, pembangunan kepercayaan, dan pembuat kebijakan yang menetapkan tingkat inflasi di lembaga makroekonomi. Dalam analisis ekonomi tentang perilaku pasar, keseimbangan teoretis permainan diasumsikan ada dalam model seperti analisis penawaran dan permintaan. Tujuan penjual adalah memaksimalkan keuntungan dalam produksi barang dan jasa.

Biaya peluang dikaitkan dengan penarikan masukan sumber daya untuk memproduksi barang. Hal ini menyebabkan kenaikan harga akibat kenaikan biaya produksi. Secara terpisah, tujuan pembeli adalah memaksimalkan utilitas. Daya beli pembeli meningkat seiring dengan penurunan harga pasar. Persaingan antara penjual dan pembeli dalam penyesuaian harga menyebabkan terjadinya keseimbangan harga. Surplus di antara penjual memaksa penurunan harga. Kekurangan di kalangan pembeli memaksa kenaikan harga. Di sini, teori makroekonomi adalah bidang ilmu ekonomi yang mempelajari penggunaan sumber daya, stabilitas harga, pertumbuhan ekonomi, dan interaksi antar negara dalam perekonomian dunia.

Sebaliknya, mikroekonomi menggambarkan perilaku ekonomi dan keputusan yang dibuat oleh pelaku ekonomi individu. Perilaku mereka mempengaruhi harga relatif yang bertindak sebagai sinyal dalam ekonomi pasar untuk memandu produksi dan konsumsi. Analisis teoritis permainan seperti ini dapat diterapkan pada pasar yang merupakan hasil interaksi strategis dibandingkan proses alami stokastik seperti pasar digital, komputasi awan, pasar energi, dan sistem crowdsourcing. Di sini, mesin komputasi menciptakan interaksi strategis melalui komunikasi dan perdagangan. Masalah inferensi statistik yang dihasilkan dalam optimasi teoretis permainan dapat mengambil manfaat dari literatur ekonometrik tentang inferensi parametrik dari interaksi strategis yang diamati.

Halpern dkk. mensurvei tema-tema utama di persimpangan teori permainan dan ilmu komputer. Kompleksitas komputasi pemodelan rasionalitas terbatas dianalisis dengan mengacu pada desain mekanisme algoritmik dalam teori permainan. Desain mekanisme tersebut dapat diterapkan dalam lelang kombinatorial untuk mekanisme pemungutan suara, lelang spektrum, slot waktu bandara, dan pengadaan industri. Di sini, “mekanisme” adalah

protokol interaksi antar pemain untuk menentukan solusi bagi masalah optimasi yang mendasarinya. Ada ketergantungan yang kompleks antara data yang diperoleh dan perilaku tertentu dalam suatu mekanisme. Secara umum, teori permainan algoritmik berbeda dari mikroekonomi dalam hal fokus pada masalah optimasi dengan solusi optimal, hasil yang tidak mungkin, jaminan perkiraan yang layak, dll. dalam jaringan mirip Internet. Narahari dkk. menulis tentang penerapan teori permainan dan desain mekanisme untuk pemecahan masalah di bidang teknik, ilmu komputer, ekonomi mikro, dan ilmu jaringan. Contoh ilustratif diberikan untuk ide-ide utama desain mekanisme seperti teori pilihan sosial, mekanisme langsung, dan mekanisme tidak langsung.

Narahari dkk. dalah monografi penelitian lain tentang teori desain mekanisme. Mekanisme optimal digambarkan sebagai arah penelitian untuk mengoptimalkan metrik kinerja seperti fungsi pembayaran permusuhan dalam pembelajaran mendalam permusuhan teoritis permainan. Mekanisme pembagian biaya diusulkan sebagai protokol untuk merancang fungsi biaya adversarial yang efisien secara komputasi dengan insentif dan anggaran. Mekanisme berulang dapat digunakan untuk mengurangi biaya penghitungan penilaian dan alokasi dalam pembelajaran mendalam adversarial teoretis permainan. Penelitian lebih lanjut tentang pembelajaran teori permainan dapat ditemukan dalam prosiding konferensi seperti Decision and Game Theory for Security (GameSec) dan Logic dan Foundations of Game and Decision Theory (LOFT).

Penambangan Data Teoritis Permainan

Fayyad dkk. mendefinisikan kerangka kerja yang disebut penemuan pengetahuan dalam database (KDD) untuk menyatukan algoritma penambangan data dengan aktivitas analisis data. KDD adalah proses desain yang memanfaatkan algoritma penambangan data untuk analisis, desain, dan penemuan pola yang berguna dalam database. KDD didefinisikan sebagai keseluruhan proses untuk menemukan pengetahuan yang berguna dari data. Penambangan data disebut sebagai penerapan teori pembelajaran mesin tertentu pada langkah-langkah tertentu dari algoritma pemecahan masalah tingkat lanjut dalam proses KDD.

Tanpa konteks yang disediakan oleh proses KDD, algoritma data mining dapat menemukan pola yang tidak berarti. Proses KDD mengambil ide dari beberapa bidang penelitian ilmu komputer seperti pembelajaran mesin, pengenalan pola, database, statistik, kecerdasan buatan, sistem pakar, visualisasi data, dan komputasi kinerja tinggi. Tujuan pemersatu adalah mengekstrak pola pengetahuan tingkat tinggi dari model data tingkat rendah dalam konteks kumpulan data kompleks yang diperoleh dari sumber data dunia nyata. Menskalakan properti matematika algoritma penambangan data ke kumpulan data besar adalah salah satu tujuan utamanya. Sistem KDD menawarkan prosedur statistik untuk pengambilan sampel dan pemodelan data, evaluasi hipotesis, dan penanganan proses kebisingan yang merugikan.

Metode KDD menggunakan lebih banyak pencarian dalam ekstraksi model untuk beroperasi dalam konteks kumpulan data besar dengan struktur data yang kaya. Tujuan KDD membedakan antara verifikasi hipotesis pembelajaran dan penemuan pola secara otonom.

Tujuan penemuan selanjutnya dikelompokkan ke dalam pemodelan prediksi dan analisis deskriptif yang terkenal. Pentingnya tugas prediksi dan deskripsi bervariasi antar aplikasi penambangan data. Metode penambangan data utama untuk mengimplementasikan aplikasi adalah seleksi, ekstraksi, klasifikasi, regresi, pengelompokan, asosiasi, peringkasan, optimasi, pengacakan, perkiraan, pemodelan ketergantungan, dan deteksi perubahan. Komponen algoritma data mining diidentifikasi sebagai fitur representasi model, kriteria evaluasi model, algoritma pembelajaran model, dan metode pencarian model.

Begoli dkk. mengusulkan proses penemuan pengetahuan untuk menganalisis data besar-besaran. Prinsip desain diberikan untuk proses pengumpulan data, organisasi sistem, dan praktik penyebaran data. Mereka mampu mengakomodasi berbagai metode analitik seperti analisis statistik, penambangan data dan pembelajaran mesin, serta visualisasi data dan analisis data eksplorasi dalam jalur analisis data. Mereka dapat menggabungkan arsitektur ringan yang mengurangi biaya, memaksimalkan kinerja, dan melacak asal untuk menyimpan, memproses, dan menganalisis data terstruktur, data semi-terstruktur, data tidak terstruktur, dan data polistruktur.

Triantaphyllou dkk. membahas metode yang efisien dan efektif untuk penambangan data dan penemuan pengetahuan (DM&KD) dalam logika matematika dan kecerdasan buatan. Metode DM&KD semacam itu dapat diterapkan dalam verifikasi formal pembelajaran mendalam permusuhan. Teknik pencarian dan algoritme pembelajaran tambahan dibahas untuk menyimpulkan fungsi Boolean monoton, aturan asosiasi, dan pembelajaran terpandu dari contoh adversarial dan contoh validasi dalam pembelajaran mesin adversarial. Grafik penolakan dapat dibuat berdasarkan contoh permusuhan berdasarkan properti menarik yang diperoleh untuk kelas positif dan negatif dalam pembelajaran permusuhan.

Subgraf klik diperoleh sebagai komponen yang terhubung dari penguraian grafik penolakan. Mereka memberikan wawasan komputasi untuk meminimalkan ukuran aturan yang disimpulkan pada contoh pelatihan. Dengan demikian, grafik penolakan juga memberikan intuisi untuk mempartisi data asli dalam masalah pembelajaran adversarial skala besar. Maimon dkk. mensurvei taksonomi metode penambangan data. Beberapa algoritma data mining disurvei untuk tugas analitik seperti pembersihan data, imputasi nilai yang hilang, ekstraksi fitur, reduksi dimensi, pemilihan fitur, metode diskritisasi, deteksi outlier, induksi aturan, pohon keputusan, jaringan Bayesian, kerangka regresi, mesin vektor dukungan, data visualisasi, aturan asosiasi, pengelompokan, klasifikasi, penambangan himpunan frekuensi, analisis tautan, optimasi multi-tujuan, jaringan saraf, pembelajaran penguatan, komputasi granular, logika fuzzy, penambangan fraktal, metode wavelet, fusi informasi, perbandingan model, ukuran ketertarikan, bahasa kueri , penambangan teks, penambangan aliran data, penambangan data spasial, penambangan data relasional, penambangan data web, penambangan data kolaboratif, dan penambangan data paralel.

L'Huillier dkk. memperkenalkan penambangan data permusuhan untuk tugas klasifikasi. Ini diterapkan pada deteksi penipuan phishing dengan pesan email berbahaya yang teknik penyaringan spamnya tidak efektif. Versi online dari mesin vektor dukungan margin tertimbang dirancang dengan teori permainan. Dalam evaluasi eksperimental, kinerjanya

lebih baik daripada algoritme klasifikasi online canggih yang beroperasi di lingkungan yang berlawanan. Teknik pemfilteran phishing sisi server dapat menggabungkan teori pembelajaran mesin yang bermusuhan untuk mengekstrak fitur yang relevan dari email phishing. Mereka kemudian dapat menggunakan algoritma data mining untuk menentukan pola tersembunyi dalam hubungan antara fitur yang diekstraksi.

Kerangka kerja penambangan data teoretis permainan memodelkan permainan sinyal antara musuh dan pengklasifikasi untuk memecahkan masalah klasifikasi permusuhan. Permainan pemberian sinyal menerapkan persyaratan keamanan untuk rasionalitas sekuensial pada fungsi imbalan yang berlawanan. Keseimbangan dan penyempurnaannya dalam permainan dinamis informasi yang tidak lengkap dapat menciptakan teori pembelajaran online seputar peristiwa tambahan yang disajikan pada pengklasifikasi online yang beroperasi dalam lingkungan yang bermusuhan. Algoritme online dan algoritme generatif harus dianggap sebagai pembelajaran mesin yang berlawanan untuk meminimalkan biaya komputasi dalam desain fungsi biaya bagi pelajar bahkan dengan mengorbankan daya prediksi yang lebih rendah untuk pembelajaran diskriminatifnya. Pengklasifikasi permusuhan yang dihasilkan memiliki aplikasi dalam phishing yang menipu dan phishing malware untuk pemfilteran phishing otomatis. Ini memiliki kinerja yang lebih baik daripada tindakan pencegahan seperti daftar hitam dan daftar putih, pemfilteran berbasis konten, otentikasi jaringan, dan enkripsi.

Bruckner dkk. memodelkan interaksi antara pembelajar dan generator dalam penyaringan spam email sebagai kompetisi pembelajaran permusuhan Stackelberg. Dengan memperhitungkan distribusi data permusuhan yang dihasilkan pada waktu penerapan serta distribusi data pelatihan yang tersedia untuk pembelajaran, permainan prediksi Stackelberg menggeneralisasi solusi pemodelan prediktif yang ada dalam konteks pemfilteran spam email. Sebuah permainan non-zero sum diusulkan di mana musuh dan pembelajar bertindak secara berurutan tanpa informasi tentang tindakan lawan. Dengan berkomitmen pada model prediktif, pelajar bertindak sebagai pemimpin permainan.

Ketika parameter model pembelajar, transformasi musuh, dan fungsi kerugian kedua pemain memenuhi kriteria matematika yang ditentukan dengan baik, permainan prediksi memiliki keseimbangan Nash unik yang dapat diselesaikan sebagai masalah optimasi. Pembelajaran teori permainan menyelesaikan program matematika bilevel dengan batasan kesetimbangan dengan penyelesaian yang diperoleh dengan metode pemrograman kuadratik sekuensial. Wang dkk. membahas integrasi teori permainan dengan penambangan data, kecerdasan buatan, dan sibernetika. Memasukkan penambangan data ke dalam teori permainan memungkinkan analisis teoretis permainan atas data kompleks dalam database dengan proses penemuan pengetahuan penambangan data. Pengetahuan yang dipelajari diwakili oleh fitur data mining seperti aturan prediksi, aturan klasifikasi, aturan asosiasi, dan aturan clustering.

Representasi seperti itu dapat mengarah pada peningkatan operasional pada pemodelan teoritis permainan dengan mendukung pengambilan keputusan berdasarkan data dalam pembelajaran mendalam yang bersifat adversarial. Metode penambangan data dapat

diterapkan dalam dinamika permainan evolusioner tidak hanya untuk pra-pemrosesan data tetapi juga pemilihan strategi dalam skenario dunia nyata yang dipilih oleh fitur penambangan data yang mengarah ke area penelitian yang disebut penambangan permainan. Selanjutnya, menurut sifat teori pemain game, jenis penambangan game didefinisikan sebagai penambangan konten game, penambangan struktur game, dan penambangan penggunaan game.

Cesa-Bianchi dkk. menulis buku tentang prediksi barisan individu dengan mengambil ide dari teori keputusan statistik, teori informasi, teori permainan, pembelajaran mesin, dan keuangan matematika. Masalah prediksi didefinisikan dalam kaitannya dengan evolusi fenomena alam dalam jangka pendek. Formalisasi prediksi diberikan dalam bentuk prediksi sekuensial menjadi realisasi proses stokastik stasioner dalam teori keputusan statistik. Di sini, sifat statistik dari proses probabilistik diperkirakan dari urutan pengamatan masa lalu untuk mendapatkan aturan prediksi atas perkiraan tersebut. Risiko aturan prediksi kemudian didefinisikan sebagai nilai yang diharapkan dari fungsi kerugian yang mengukur perbedaan antara nilai prediksi dan hasil sebenarnya.

Kinerja prediktor kemudian diukur berdasarkan kerugian kumulatif yang terakumulasi dalam banyak putaran prediksi. Tanpa proses stokastik yang mendasari prediktor, tidak ada dasar untuk membandingkan kinerja prediktor. Garis dasar tersebut kemudian dimodelkan dari sekelompok model yang disebut peramal referensi yang menggabungkan keahlian domain untuk memberikan saran mengenai hasil selanjutnya. Pakar domain juga dapat memasukkan model kotak hitam dengan kekuatan komputasi yang tidak diketahui dan akses ke sumber informasi sampingan pribadi. Kelas pakar dapat menjadi model statistik yang dibangun berdasarkan fenomena alam. Perbedaan antara kerugian kumulatif seorang ahli dan seorang prediktor didefinisikan sebagai penyesalan. Strategi pembelajaran mesin dirancang untuk meminimalkan penyesalan semua pakar di kelas konsep atas pembelajaran permusuhan.

Ketangguhan permusuhan juga dapat didefinisikan dalam istilah minimalisasi penyesalan. Algoritme prediksi pengacakan kemudian dapat dirancang untuk memainkan permainan berulang untuk memprediksi keputusan gabungan berurutan yang mengompresi urutan prediksi menjadi keputusan gabungan. Di sini, teori informasi dapat dimanfaatkan untuk melakukan kompresi data dengan mengacu pada fungsi kerugian tertentu. Distribusi probabilitas ditentukan pada kumpulan hasil yang mungkin menggunakan statistik Bayesian untuk estimasi kemungkinan maksimum dalam pengenalan pola online. Lingkungan stokastik yang menghasilkan rangkaian prediksi dapat dipelajari dengan teori permainan. Teorema minimax dalam teori permainan dapat memberikan batasan pada kinerja algoritma prediksi sekuensial. Teorema minimax yang digeneralisasi dapat digunakan untuk menentukan jaminan kinerja untuk algoritma pembelajaran mendalam yang bermusuhan tanpa mengasumsikan distribusi probabilitas bentuk tertutup yang mendasari distribusi data yang bermusuhan. Kemudian strategi minimalisasi penyesalan dalam pembelajaran adversarial teori permainan menginduksi dinamika serangan-pertahanan yang mengarah pada beberapa pengertian keseimbangan teori permainan.

Rezek dkk. menetapkan kosakata umum untuk inferensi statistik dalam teori permainan dan pembelajaran mesin. Analogi terbentuk antara respon terbaik dalam permainan fiktif dan metode inferensi Bayesian. Aturan pembaruan untuk permainan variasi fiktif diusulkan untuk algoritma pembelajaran variasional. Mereka menunjukkan sifat konvergensi yang lebih baik dalam model grafis yang sangat terhubung dan memiliki aplikasi untuk mengelompokkan distribusi campuran. Tujuan pengoptimalan diselesaikan sehubungan dengan distribusi gabungan fitur pembelajaran yang dipilih oleh semua pemain dalam game.

Kleinberg dkk. menyajikan kerangka optimasi yang kuat untuk evaluasi operasi penambangan data dalam pengambilan keputusan seperti asosiasi dan pengelompokan dengan fungsi utilitas dalam teori permainan. Ketertarikan dan nilai dari pola penambangan data ditentukan oleh sejauh mana pola tersebut dapat digunakan untuk proses pengambilan keputusan berbasis data di suatu perusahaan guna meningkatkan kegunaan/nilai teoritis permainan dari keputusan yang dibuat oleh perusahaan selama interaksinya dengan pihak lain. agen di pasar seperti pelanggan, pemasok, karyawan, pesaing, pemerintah, dll.

Pandangan utilitarian terhadap penambangan data perlu menjawab beberapa pertanyaan penelitian dalam optimasi kombinatorial, pemrograman linier, dan teori permainan. Permainan matriks diusulkan untuk menyelesaikan masalah clustering dengan algoritma aproksimasi. Kompleksitas komputasinya dapat ditingkatkan secara berulang melalui metode pengambilan sampel data dan pembelajaran serakah. Analisis sensitivitas masalah optimasi dilakukan untuk mengusulkan ukuran ketertarikan baru untuk pola data mining. Kemudian dikembangkan menjadi teori prediksi nilai operasi data mining dengan metode penemuan pengetahuan dan alat kecerdasan buatan yang mampu secara eksplisit memperhitungkan maksud dan tujuan tugas data mining.

Freitas dkk. survei algoritma evolusioner (EA) yang bertindak sebagai algoritma pencarian stokastik dalam penambangan data. Algoritme genetik (GA) dan pemrograman genetik (GP) adalah kelas EA yang populer yang disajikan sebagai teknik pencarian adaptif yang kuat dalam menyelesaikan tugas penambangan data seperti penemuan aturan klasifikasi, konstruksi dan seleksi atribut, dan pengelompokan. EA multi-objektif dapat menemukan solusi optimal Pareto dalam beberapa tugas penambangan data yang dapat diterapkan untuk menyusun contoh permusuhan dalam penelitian kami tentang algoritma pembelajaran permusuhan teoritis permainan.

EA juga dapat mengatasi interaksi atribut yang beragam dengan lebih baik dalam tugas penambangan data yang berlawanan. Fungsi kerugian khusus yang diusulkan untuk setiap jenis musuh dapat dimodelkan sebagai fungsi kesesuaian yang mengevaluasi solusi kandidat untuk contoh permusuhan berdasarkan berbagai kriteria kualitas dalam EA. Pembelajaran mesin berbasis instans dapat menerima representasi data permusuhan untuk penambangan data dengan EA secara sinergis untuk merancang representasi individu, fungsi kebugaran, dan operator genetika khusus untuk tugas penambangan data permusuhan yang sedang diselesaikan. Dengan secara otomatis menemukan program komputer dengan dokter, EA

dapat digunakan untuk induksi algoritma yang melampaui induksi aturan dalam pembelajaran mendalam yang bermusuhan.

Ficici dkk. membedakan antara algoritma evolusi dan algoritma ko-evolusioner dalam hal interaksi antara entitas yang berevolusi bersama. Teori permainan digunakan untuk menggambarkan interaksi tersebut dengan mengasumsikan ko-evolusi sebagai metode optimasi yang mengarah pada perluasan model rantai Markov untuk algoritma evolusi. Teori pembelajaran komputasi (COLT) dapat digunakan untuk menganalisis permainan kompetitif untuk membangun dinamika dan batasan algoritma ko-evolusi. Ficici dkk. menggunakan teori permainan evolusi untuk pemilihan fitur dalam algoritma penambangan data ko-evolusi.

Permainan jumlah variabel diusulkan untuk peringkat linier, pemilihan Boltzmann, dan pemilihan fitur turnamen. Seleksi Boltzmann menyatu ke dalam keseimbangan Nash polimorfik menurut penarik titik dari teori chaos dalam fisika teoretis dan sistem dinamik. Ekuilibria Nash polimorfik adalah ekuilibria Nash untuk permainan strategi campuran yang mengekspresikan populasi data polimorfik. Algoritma ko-evolusi dipahami sebagai metode pencarian atau pemecah masalah dan model sistem dinamis. Ficici dkk. meneliti metode pemilihan fitur seperti kebugaran-proporsional, peringkat linier, pemotongan, dan pemilihan ES dalam konteks ko-evolusi dua populasi dengan pembelajaran teoretis permainan. Metode seleksi menambahkan wilayah ruang fase yang mengarah pada dinamika siklik pada penarik non-Nash.

Herbert dkk. menerapkan teknik teori permainan untuk menilai kualitas optimalisasi kelompok pembelajaran kompetitif dengan peta pengorganisasian mandiri (SOM). SOM menawarkan model ketahanan yang fleksibel untuk pengelompokan dengan beberapa aspek yang dapat dikonfigurasi dalam banyak aplikasi berbeda. Hal ini dapat memanfaatkan struktur data yang dinamis dan adaptif untuk menentukan pembaruan neuron dalam pembelajaran kompetitif dengan mengacu pada beberapa ukuran kinerja dan kriteria seleksi dalam pembelajaran mesin. Di sini, pembelajaran teori permainan digunakan untuk meningkatkan kualitas pembaruan tidak hanya pada satu neuron tetapi juga seluruh cluster neuron dengan algoritma pelatihan yang disebut GTSOM.

Garg dkk. menerapkan teknik teori permainan untuk pengelompokan fitur. Fitur dipandang sebagai pemain rasional dalam permainan koalisi dimana koalisi adalah kelompoknya. Cluster kemudian dibentuk untuk memaksimalkan hasil individu pada konsep solusi yang disebut partisi stabil Nash (NSP). NSP diselesaikan dengan program integer linear (ILP). ILP dimodifikasi menjadi pendekatan pengelompokan hierarki untuk menemukan cluster pada sejumlah besar fitur. Dengan demikian, teori permainan digunakan dalam pemilihan fitur untuk membedakan antara fitur yang relevan dan tidak relevan serta fitur yang dapat disubstitusi dan saling melengkapi.

Shah dkk. model permainan survei dalam pelestarian privasi, keamanan jaringan, deteksi intrusi, dan optimalisasi sumber daya. Teori permainan merupakan salah satu pendekatan penambangan data yang menjaga privasi (PPDM). PPDM telah memanfaatkan aturan asosiasi untuk mencapai algoritma penambangan aturan asosiasi terdistribusi (PPRADM) yang menjaga privasi. Teori permainan juga dapat digunakan untuk merancang

trade-off antara utilitas data dan pelestarian privasi dengan model permainan berurutan. Permainan privasi dapat dirancang dari Teori Permainan Koperasi untuk menciptakan Privasi Koperasi dalam koalisi.

Teori permainan dapat digunakan dalam analisis serangan jaringan seperti serangan browser, serangan penolakan layanan (DDoS), serangan worm, dan serangan malware. Model permainan honeypot Bayesian telah diusulkan untuk memecahkan masalah yang disebabkan oleh serangan penolakan layanan terdistribusi. Model Stackelberg telah diusulkan untuk masalah pengerasan jaringan dimana pembela secara optimal menambahkan honeypots dalam jaringan untuk mendeteksi penyerang. Model permainan dinamis telah diusulkan untuk sistem deteksi intrusi (IDS) di jaringan nirkabel ad hoc. Optimasi IDS dapat dikategorikan menjadi optimasi alokasi sumber daya, optimasi konfigurasi IDS, dan optimasi penanggulangan.

Masalah optimasi alokasi sumber daya berkaitan dengan optimalisasi pengambilan sampel tautan jaringan, pembagian sumber daya antar node, strategi pertahanan cluster dalam jaringan sensor, dll. Optimasi konfigurasi IDS berkaitan dengan optimalisasi sensitivitas IDS, kemampuan bertahan hidup, dan mitigasi serangan dalam jaringan sensor nirkabel. Optimalisasi penanggulangan berkaitan dengan optimalisasi waktu tidak tersedianya node jaringan, penghitungan respons optimal dalam serangan multi-tahap, penghitungan penanggulangan optimal dalam jaringan sensor nirkabel, dll.

Dia dkk. menggunakan teori permainan untuk mekanisme lelang kata kunci dalam pencarian bersponsor untuk memilih iklan dalam monetisasi mesin pencari. Mekanisme lelang dirumuskan dalam kerangka optimasi bilevel yang diselesaikan dengan pembelajaran mesin teori permainan. Pemilihan iklan ditentukan oleh peringkat dan harga iklan untuk lelang kata kunci. Model Markov pada data historis menggambarkan perubahan tawaran pengiklan sebagai respons terhadap mekanisme lelang. Indeks kinerja utama (KPI) yang memenuhi properti Markov mencakup sinyal ke mesin pencari seperti jumlah tayangan, jumlah klik, dan rata-rata biaya per klik.

Setiap pengiklan diasumsikan tidak memiliki detail apa pun tentang perilaku penawaran pengiklan yang tersisa untuk kata kunci dalam mekanisme lelang. Selain itu, setiap pengiklan tidak memiliki pengetahuan tentang internal mekanisme lelang. Jadi model perilaku pengiklan yang bergantung pada mekanisme dapat dibangun dari log lelang historis. Model Markov kemudian mampu memprediksi urutan penawaran di masa depan. Mekanisme lelang kemudian dirancang secara empiris untuk memaksimalkan pendapatan berdasarkan urutan penawaran yang diprediksi. Model pendapatan empiris menyatu ketika jangka waktu prediksi mendekati tak terhingga. Algoritme pemrograman genetik mengoptimalkan model pendapatan empiris. Dengan demikian, pembelajaran mesin teoretis permainan dapat diterapkan dalam perdagangan elektronik dan kecerdasan buatan. Algoritme pemrograman genetik dapat menangani hubungan fungsional non-linier yang kompleks dalam rangkaian prediksi untuk mengambil strategi penawaran respons terbaik.

Narayanam dkk. merancang algoritma pembelajaran teoretis permainan untuk mendeteksi komunitas yang tidak tumpang tindih dalam jaringan sosial yang dibentuk oleh

tindakan individu rasional dalam jaringan Internet. Hasilnya sebanding dengan teknologi terkini dalam pengelompokan grafik, misalnya karena pendekatan partisi bertingkat pada pengelompokan. Kegunaan suatu node dalam komunitas didefinisikan sebagai jumlah tetangga dari node tersebut dalam komunitasnya dan sebagian kecil dari tetangga dalam komunitasnya yang terhubung sendiri.

Modularitas dan cakupan node adalah ukuran kinerja dalam eksperimen pengelompokan untuk menemukan komunitas di jaringan sosial. Bulo dkk. mengekstrak kelompok pengelompokan hipergraf dengan menggunakan teori permainan untuk memformalkan gagasan cluster. Masalah pengelompokan yang menggunakan kesamaan tingkat tinggi antar objek disebut masalah pengelompokan hipergraf. Sebuah permainan clustering multipemain non-kooperatif dirancang untuk menemukan kualitas clustering sesuai dengan konsep solusi keseimbangan teoritis permainan. Menemukan keseimbangan permainan pengelompokan terbukti setara dengan mengoptimalkan fungsi polinomial secara lokal dengan batasan linier. Dinamika waktu diskrit digunakan untuk mengoptimalkan fungsi polinomial. Hasilnya dibandingkan dengan metode ekspansi klik pada hipergraf berbobot tepi. Cluster yang dihasilkan menunjukkan ketahanan terhadap outlier.

Freund dkk. membahas hubungan antara teori permainan dan pembelajaran online. Masalah prediksi acak dalam pembelajaran mesin online didefinisikan sebagai model pembelajaran di mana agen memprediksi klasifikasi serangkaian item sambil meminimalkan jumlah kesalahan prediksi. Algoritme peningkatan diusulkan untuk menggabungkan model pembelajaran yang diperoleh dari beberapa proses di beberapa distribusi data. Ini menggabungkan beberapa hipotesis yang dipilih menjadi hipotesis akhir dengan tingkat kesalahan yang sangat kecil. Kesalahan generalisasi hipotesis akhir dapat dibatasi dengan mengacu pada teori VC teori pembelajaran komputasi.

Musuh yang Sensitif terhadap Biaya

Nelson dkk. mempelajari bagaimana musuh dapat menanyakan pengklasifikasi secara efisien. Contoh permusuhan yang tidak terdeteksi dibuat dengan biaya minimum bagi musuh menggunakan jumlah kueri polinomial di ruang fitur pelatihan. Dengan demikian, musuh yang sensitif terhadap biaya dapat menemukan titik buta dari sebuah detektor dengan mengamati respons permintaan keanggotaan dari detektor terhadap label negatif untuk membangun contoh permusuhan berbiaya rendah yang memiliki dampak maksimal pada kinerja yang diinginkan dari detektor. Masalah dalam menemukan contoh negatif berbiaya rendah dengan sedikit pertanyaan ini disebut masalah penghindaran mendekati optimal.

Pengklasifikasi yang ditargetkan disebut pengklasifikasi penginduksi konveks. Ini mencakup pengklasifikasi linier dan detektor anomali yang mempelajari batasan keputusan hipersfer. Tidak perlu merekayasa balik batas keputusan pengklasifikasi. Tujuan yang berlawanan dari pengoptimalan berbasis kueri sebanding tetapi tidak serupa dengan bidang penelitian pembelajaran aktif. Gagasan musuh mengenai kegunaan untuk menciptakan contoh-contoh permusuhan diwakili oleh fungsi biaya permusuhan. Lanckriet dkk. menganalisis kemungkinan kesalahan klasifikasi dari klasifikasi yang benar dari titik data masa depan dalam pengaturan kasus terburuk untuk pengklasifikasi.

Masalah minimax yang dihasilkan diinterpretasikan secara geometris sebagai minimalisasi jarak Mahalanobis maksimum antara dua kelas dalam masalah klasifikasi biner yang dioptimalkan oleh program kuadrat. Ketahanan pengklasifikasi ditentukan berdasarkan kesalahan estimasi rata-rata dan kovarians kelas. Hal ini ditemukan kompetitif dengan pengklasifikasi non-linier seperti mesin vektor dukungan. Ini adalah pendekatan diskriminatif untuk mengukur kekuatan pengklasifikasi yang bersifat permusuhan. Hal ini dapat dibandingkan dengan pendekatan generatif yang membuat asumsi distribusi tentang kepadatan bersyarat kelas dalam data adversarial untuk memperkirakan dan mengontrol probabilitas yang relevan.

Asif dkk. membahas hukuman yang bergantung pada aplikasi untuk kesalahan antara label kelas yang diprediksi dan sebenarnya dalam pengklasifikasi yang kuat. Biaya kesalahan dirumuskan sebagai masalah optimasi cembung pada kerugian sensitif biaya non-cembung. Pendekatan terhadap ketahanan permusuhan ini kontras dengan minimalisasi risiko empiris pada kerugian pengganti cembung yang dapat dilakukan. Namun, perbedaan statistik antara kerugian aktual dan pengganti cembungnya dapat menyebabkan ketidaksesuaian yang signifikan secara statistik antara estimasi parameter optimal berdasarkan fungsi kerugian pengganti dan tujuan kinerja awal.

Hukuman atas kesalahan direpresentasikan sebagai matriks biaya kebingungan untuk tugas klasifikasi. Berbeda dengan metode pembobotan ulang dan kerugian akibat kesalahan tertentu, tujuan pembelajaran mesin yang diawasi adalah meminimalkan biaya kesalahan klasifikasi yang diperkirakan. Konstruksi pengklasifikasi dibingkai sebagai permainan melawan evaluator yang bermusuhan. Estimasi parameter dalam pengklasifikasi dinyatakan sebagai solusi terhadap parameter hasil permainan zero-sum yang secara efisien ditemukan oleh pemrograman linier. Batasan kinerja diberikan pada kesalahan generalisasi untuk menunjukkan manfaat empiris dari pendekatan yang diusulkan untuk konstruksi pengklasifikasi.

Pembelajaran sensitif biaya yang diusulkan memiliki fungsi kerugian yang merugikan yang bergantung pada kelas aktual dan kelas yang diprediksi. Ini lebih umum daripada kerugian nol-satu yang digunakan dalam klasifikasi biner untuk mempelajari prediksi kelas terbaik. Namun memperkirakan distribusi kelas bersyarat untuk pembelajaran yang sensitif terhadap biaya memerlukan lebih banyak data pelatihan yang ditemukan di sumber data atau dihasilkan secara sintetis. Pembelajar yang sensitif terhadap biaya dapat dimanfaatkan selama pelatihan atau waktu prediksi untuk secara berulang menimbang ulang kumpulan data pelatihan yang tersedia sehingga pengklasifikasi sensitif terhadap kesalahan yang merugikan dalam tugas prediksi kelas jamak. Biaya kebingungan juga dapat dimasukkan ke dalam kriteria klasifikasi. Teknik peningkatan yang sensitif terhadap biaya dapat menggabungkan beberapa pembelajar lemah yang sensitif terhadap biaya untuk menghasilkan pembelajar yang kuat.

Di sini, perspektif permusuhan terhadap pengklasifikasi yang sensitif terhadap biaya memperkenalkan estimasi statistik dan pengambilan keputusan dalam kondisi ketidakpastian ke dalam konstruksi pengklasifikasi. Prosedur statistik untuk pengambilan keputusan

berdasarkan data mencakup model maximin Wald, permainan permusuhan sekuensial, optimalisasi penyesalan keputusan minimax Savage, perkiraan statistik dalam ketidakpastian yang meminimalkan risiko kasus terburuk, dan pemodelan entropi maksimum untuk distribusi keluarga eksponensial dalam permusuhan.

kehilangan. Formulasi pembelajaran mesin adversarial seperti itu tahan terhadap pergeseran adversarial yang bertindak sebagai kendala pada statistik momen kumpulan fitur pelatihan dan ketidakpastian dalam estimasi fungsi kerugian yang sensitif terhadap biaya pada distribusi label bersyarat. Matriks biaya berparameter menentukan hasil permainan setiap pemain. Kumpulan fitur pelatihan dapat menggabungkan metode kernel untuk mempertimbangkan ruang fitur yang lebih kaya yang mendekati data pelatihan.

Untuk menjelaskan perbedaan dalam pembuatan data dan prosedur transmisi, De Silva et al. mempertimbangkan karakteristik kerentanan data uji dalam desain tindakan penanggulangan pengklasifikasi permusuhan. Jadi struktur biaya serangan dapat dipelajari dalam lingkungan protokol sensor-ke-keputusan seperti dalam sistem Internet of Things. Ini juga dapat digunakan untuk melakukan analisis kerentanan terhadap sistem pembelajaran mesin yang diterapkan di dunia nyata. Kontribusi utama dari kerangka pembelajaran adversarial sadar biaya (CAL) yang diusulkan adalah operator proyeksi untuk mengurangi dampak pemalsuan. Ini memproyeksikan contoh pengujian yang dipalsukan ke ruang vektor fitur yang sah dengan mengacu pada fungsi biaya serangan yang bertindak sebagai metrik jarak.

Pendekatan CAL dievaluasi pada model campuran Gaussian (GMM) dengan analisis komponen utama (PCA) dan pengklasifikasi jaringan saraf dalam (DNN). Parameter GMM diestimasi menggunakan algoritma ekspektasi-maksimisasi (EM) pada distribusi bersyarat kelas. Fungsi biaya didefinisikan sebagai fungsi norma kuadrat selain fungsi biaya norma L1. Serangan permusuhan diasumsikan sebagai serangan kotak putih dan serangan kotak abu-abu di mana arsitektur jaringan saraf dalam diketahui oleh musuh tetapi parameter yang dipelajari mungkin diketahui atau tidak.

Rios Insua dkk. meninjau keadaan seni dalam klasifikasi permusuhan dari pertimbangan kerangka teori permainan yang kontras dengan analisis risiko permusuhan. Analisis risiko permusuhan yang diusulkan adalah masalah analisis keputusan Bayesian. Hal ini tidak mengasumsikan bahwa agen teori permainan berbagi informasi tentang keyakinan dan preferensi mereka berdasarkan hipotesis pengetahuan umum dalam kerangka teori permainan. Strategi ketahanan permusuhan bergantung pada apakah pengklasifikasi generatif atau diskriminatif sebagai model dasar. Simulasi Monte Carlo (MC) memecahkan masalah serangan yang optimal. Teknik perkiraan perhitungan Bayesian (ABC) menghasilkan distribusi data yang berlawanan.

Masalah klasifikasi biner dievaluasi dalam pengaturan serangan penghindaran dan serangan pelanggaran integritas. Di sini, kerangka teori permainan dapat membuat kesimpulan real-time tentang proses pengambilan keputusan lawan pada waktu operasi. Analisis risiko permusuhan cocok untuk aplikasi yang memiliki hambatan komputasi dengan kemungkinan perubahan perilaku musuh yang dimasukkan ke dalam pelatihan ulang

pengklasifikasi. Klasifikasi permusuhan yang diusulkan memiliki aplikasi keamanan siber dalam proses otomatisasi yang ditemukan dalam deteksi spam, mengemudi otonom, deteksi penipuan, deteksi phishing, pemfilteran konten, penyaringan kargo, kebijakan prediktif, dan terorisme.

Fawzi dkk. menganalisis ketahanan pengklasifikasi terhadap gangguan permusuhan untuk mendapatkan batas atas ukuran ketahanan permusuhan pada kesulitan tugas klasifikasi. Ukuran ketahanan bergantung pada ukuran kemampuan membedakan antar kelas. Ketidakstabilan permusuhan disebabkan oleh rendahnya fleksibilitas pengklasifikasi dibandingkan dengan kesulitan tugas klasifikasi. Perbedaan dibuat antara ketahanan pengklasifikasi terhadap gangguan acak dan ketahanannya terhadap gangguan permusuhan. Dalam tugas klasifikasi dunia nyata, konsep ketahanan adversarial yang lemah berhubungan dengan informasi parsial tentang tugas klasifikasi, sedangkan konsep yang kuat menangkap esensi kelompok klasifikasi tetap seperti fungsi linier sepotong-sepotong untuk tugas klasifikasi.

Kelompok pengklasifikasi non-linier yang lebih fleksibel dan algoritma pelatihan yang lebih baik ditemukan untuk mencapai ketahanan yang lebih baik. Evaluasi eksperimental menunjukkan bahwa peningkatan kedalaman jaringan saraf membantu meningkatkan ketahanan lawannya, namun menambahkan lapisan ke jaringan yang sudah dalam hanya sedikit mengubah ketahanan tersebut. Biggio dkk. meningkatkan ketahanan pengklasifikasi dengan strategi penyembunyian informasi yang memperkenalkan keacakan dalam fungsi keputusan. Ini digunakan dalam arsitektur sistem pengklasifikasi ganda. Formulasi teoritis permainan antara pengklasifikasi dan musuh diusulkan untuk menciptakan pengklasifikasi sadar-musuh yang berkinerja lebih baik. Kurangnya informasi mengenai batasan keputusan yang tepat menyebabkan pihak lawan mengambil pilihan yang terlalu konservatif atau terlalu berisiko dalam memutuskan manipulasi permusuhan untuk pola yang jahat.

Dengan demikian, pengklasifikasi bisa mendapatkan keuntungan dengan meningkatkan ketidakpastian musuh. Namun, pengacakan yang berlebihan juga dapat menyebabkan penurunan kinerja pengklasifikasi yang dipilih. Pertukaran antara strategi pengacakan ini dianalisis dengan permainan strategi yang berulang-ulang untuk memungkinkan pengklasifikasi berlatih kembali sesuai dengan strategi yang dipilih oleh lawan. Schmidt dkk. menganalisis kompleksitas sampel pembelajaran yang kuat dalam pengklasifikasi canggih yang tunduk pada gangguan permusuhan. Hal ini penting dalam menganalisis sifat ketahanan sistem pembelajaran yang diterapkan di lingkungan yang kritis terhadap keselamatan dan keamanan. Kompleksitas sampel dari generalisasi pengklasifikasi jinak standar dibandingkan dengan kompleksitas sampel dari generalisasi kuat yang berlawanan dalam model distribusi tertentu seperti model Gaussian dan model Bernoulli.

Batas atas dan bawah diberikan sebagai jaminan sampel terbatas untuk kompleksitas sampel jika terjadi pergeseran distribusi dalam kasus terburuk. Keberadaan musuh dianggap secara signifikan meningkatkan kerugian hipotesis apa pun dalam kelas hipotesis untuk desain pengklasifikasi. Bentuk ketahanan yang tidak terlalu bermusuhan seperti mempelajari pengklasifikasi yang kuat dalam lingkungan yang tidak berbahaya dapat diterapkan pada

masalah dalam pembelajaran transfer dan adaptasi domain. Analisis kompleksitas sampel dapat diperluas dengan gangguan yang lebih terbatas dan berdimensi lebih rendah dalam rangkaian gangguan yang berbeda. Properti ketahanan lebih lanjut dipahami dalam kaitannya dengan properti distribusi adversarial yang membuat generalisasi yang kuat menjadi sulit atau mudah dalam kelas model yang diminati.

Miller dkk. meneliti pembelajaran aktif permusuhan untuk menemukan strategi alokasi sumber daya terbatas dalam pembelajaran mesin untuk memberi label pada data pelatihan dan ekstraksi fitur. Teknik pembelajaran aktif disajikan dalam konteks permusuhan yang mengharuskan pelabelan konten baru secara akurat dan tepat waktu untuk menjaga kinerja deteksi dalam aplikasi seperti platform anti-phishing dan deteksi iklan berbahaya. Dalam pembelajaran aktif, algoritme pembelajaran secara aktif melibatkan oracle untuk meminta informasi guna memberi label pada kumpulan data pelatihan. Pelajar menggunakan strategi kueri untuk memilih contoh yang akan diberi label. Itu dapat memilih contoh yang bahkan tidak muncul dalam kumpulan data.

Jadi pembelajar terbatas pada kumpulan kejadian yang diamati namun tidak diberi label dimana pelabelan oleh manusia memerlukan biaya yang mahal. Strategi pemilihan pembelajaran aktif dapat digunakan untuk memprioritaskan pelabelan manusia dengan algoritma prioritas yang disesuaikan yang memprediksi tingkat sumber daya manusia yang optimal. Ramalan yang berisik menyediakan lingkungan permusuhan yang lemah untuk pembelajaran aktif. Musuh berupaya menyamarkan kejadian berbahaya secara efektif tanpa mengeluarkan biaya. Pembela HAM mencoba mengidentifikasi kejadian berbahaya secara efisien dengan mengukur fitur dengan biaya rendah. Pembelajaran aktif seperti itu dapat dimodelkan dalam kerangka pembelajaran adversarial teoretis permainan. Biaya pengukuran fitur selama pelatihan atau pengujian kemudian menjadi subjek pembelajaran aktif.

Di sini, oracle kueri seharusnya memberi label pada instance dan menilai fitur untuk meningkatkan kinerja sistem pembelajaran. Dalam serangan whitebox, musuh diasumsikan mampu memperkirakan fungsi keputusan dengan melakukan probing berulang kali pada setiap putaran pembelajaran aktif dengan presisi yang berubah-ubah. Musuh memiliki pengetahuan tentang proses stokastik yang menghasilkan data seputar label berbahaya dan berbahaya, namun tidak mengetahui realisasi sebenarnya dari data pelatihan. Pada setiap putaran pembelajaran aktif, musuh memasukkan satu contoh saja ke dalam set pelatihan untuk mencerminkan non-stasioneritas dalam proses pembelajaran dengan memasukkan contoh-contoh terpencil dari label yang diberikan.

Jenis Oracle dikategorikan sebagai oracle ahli yang menyediakan pelabelan sangat akurat yang didukung oleh pakar teknis yang mahal, oracle berisik yang mewakili informasi crowdsourced dengan akurasi yang bervariasi berdasarkan fungsi contoh, dan oracle jahat yang menggunakan strategi permusuhan untuk memberi label yang salah pada sampel target tertentu. Pembuatan sampel permusuhan berperan dalam merancang ramalan jahat yang menghasilkan contoh umpan untuk mengurangi kualitas pembelajaran aktif. Miller dkk. meninjau pembelajaran aktif permusuhan dengan strategi pemilihan sampel campuran. Pembelajaran mesin yang peka terhadap keamanan memiliki aplikasi dalam sistem deteksi

intrusi jaringan (NIDS); otentikasi biometrik; spam email; pengenalan gambar, karakter, dan ucapan; dan klasifikasi dokumen.

Taksonomi serangan waktu pelatihan disebut tampering. Taksonomi serangan waktu pengujian/penggunaan disebut menggagalkan. Dalam serangan gangguan, tujuan permusuhan diberikan sebagai pelabelan prediksi oracle yang salah pada contoh yang diberi label dengan benar dan manipulasi fitur yang dipilih untuk membiaskan proses pembelajaran dengan optimalisasi kotak hitam. Ramalan yang berisik ini memiliki akurasi yang bervariasi ketika kebenaran dasarnya tidak diketahui. Uncertainty sampling dan Max- Expected Utility dipilih sebagai strategi kriteria pemilihan sampel dalam pembelajaran aktif. Dalam pengambilan sampel ketidakpastian, sampel yang paling tidak pasti menerima skor paling tidak pasti. Kriteria pemilihan sampel acak bertindak sebagai dasar.

Strategi kueri lainnya mencakup metode pengambilan sampel berbasis kepadatan, kueri per komite, dan pengurangan varians. Strategi campuran dapat dirancang untuk menemukan kelas yang tidak diketahui dalam data pelatihan. Dalam aplikasi keamanan, kelas yang tidak diketahui tersebut diharapkan terjadi karena data menunjukkan penyimpangan yang berlawanan dengan sifat non-stasioneritas dari distribusi data yang berubah seiring waktu. Teknik pembelajaran adversarial mengatasi penyimpangan data dengan teknik pengambilan sampel seperti pengambilan sampel ketidakpastian dengan pengacakan, menangani label berisik yang dihasilkan oleh non-adaptif, teori pengaturan pergeseran kovariat untuk musuh adaptif.

Di sini, pembelajaran aktif dalam konteks permusuhan harus mencakup eksperimen seputar kinerja sistem pembelajaran sehubungan dengan penyimpangan konsep waktu nyata, penurunan kinerja dari waktu ke waktu, pengembalian upaya manusia untuk memberi label pada data, mengatasi serangan dan pertahanan terhadap musuh jahat yang memberi label pada data, fitur biaya ekstraksi dalam analisis statis dan dinamis malware, kinerja strategi kueri untuk mempelajari suatu konsep dan bereaksi terhadap penyimpangannya, dan ketahanan strategi kueri terhadap manipulasi jahat dari oracle manusia.

Penelitian kami terhadap teori optimasi terbatas dalam fungsi tujuan untuk pembelajaran adversarial didorong oleh kemampuan dan kontrol musuh pada data pelatihan dan data validasi dengan mempertimbangkan batasan spesifik aplikasi, efek pada kelas sebelumnya, fraksi sampel, dan fitur yang dimanipulasi oleh musuh. . Tergantung pada tujuan, pengetahuan, dan kemampuan musuh, kendala ini juga diklasifikasikan berdasarkan pengaruh serangan, pelanggaran keamanan, dan kekhususan serangan. Masalah optimasi terbatas pada arsitektur dangkal cenderung menghasilkan algoritma komputasi yang sulit untuk estimasi kelas dan inferensi dalam pembelajaran yang diawasi.

Fungsi biaya adversarial yang diusulkan dan prosedur pelatihan adversarial memerlukan kebutuhan akan arsitektur pembelajaran mendalam dalam metode statistik yang memecahkan masalah optimasi dalam fungsi hasil adversarial. Pendekatan keamanan siber untuk mengambil contoh masalah dalam pembelajaran adversarial kemudian dapat berfokus pada peningkatan ketahanan pada metode rantai Markov dan permainan Bayesian

Stackelberg. Dalam konteks ini, kita dapat memperoleh model klasifikasi yang kuat dalam kerangka pembelajaran adversarial.

Algoritme penelusuran evolusioner dari penelitian kami dapat diperluas ke dalam proses pengambilan keputusan Markov dan automata seluler dengan perluasan prosedur pengoptimalan lokal yang memaksimalkan fungsi hasil yang merugikan. Di sini, jaringan kepadatan campuran dapat mengekspresikan distribusi data bersyarat pada variabel laten dan label kelas dalam data pelatihan dan data permusuhan. Kami juga dapat mengukur perbedaan informasi antara representasi minimal data pelatihan dan penyematan fitur data adversarial dengan fungsi biaya adversarial berbasis pembelajaran metrik yang mendalam.

Kami juga dapat menerapkan distribusi sebelumnya pada faktor laten untuk pembuatan data yang koheren dalam pembelajaran yang diawasi. Keandalan pembelajaran mesin tersebut dalam penerapannya dapat disimulasikan dengan pengoptimalan komputasi dan masalah inferensi statistik dalam analisis tingkat lanjut dengan musuh teoretis permainan dan kontrol sistem dinamis dalam pengoptimalan kotak hitam pembelajaran mendalam. Kami kemudian dapat menganalisis dekomposisi bias-varians dalam fungsi imbalan yang merugikan untuk mendapatkan batasan utilitas untuk pembelajaran mendalam dalam kerangka kerja yang terikat kesalahan untuk keamanan siber.

Kami mempelajari fungsi biaya adversarial yang ada sehubungan dengan batasan ketahanan dan anggaran privasi dalam model pembelajaran representasi renggang untuk pembelajaran adversarial. Untuk mendapatkan jaminan yang dapat diandalkan pada keamanan jaringan saraf, kami melakukan evaluasi keamanan pembelajaran mesin berbasis data di titik persimpangan antara pengujian perangkat lunak, verifikasi formal, kecerdasan buatan yang kuat, dan pembelajaran mesin yang dapat ditafsirkan. Di sini, kompleksitas model (atau disebut kesalahan generalisasi) dapat didefinisikan sebagai perbedaan antara kesalahan di luar sampel dan kesalahan dalam sampel dalam verifikasi formal yang diterapkan pada metode pemrosesan informasi siber dalam masalah klasifikasi pembelajaran adversarial dan masalah optimasi.

Kerangka kerja pembelajaran adversarial teoritis permainan kami dapat mengotomatiskan deteksi, klasifikasi, pembuatan, dan optimalisasi pembelajaran mesin yang dapat dipercaya di web dan aplikasi seluler. Representasi generatif dalam manipulasi permusuhan kami mampu mengukur ancaman keamanan yang mengeksploitasi pengukuran aktif dan pasif dalam domain aplikasi data besar. Kami memodelkan aktivitas jahat musuh dalam fungsi tujuan pengoptimalan teoretis game. Kemudian solusi pembelajaran mendalam yang seimbang mampu mengidentifikasi kebocoran informasi pribadi yang tersedia di platform AI. Dengan memodelkan kebocoran informasi sebagai fungsi kerugian dalam pembelajaran mendalam, kita dapat merumuskan pengaturan permusuhan dalam kerangka pembelajaran teori permainan.

Kita dapat memasukkan skenario serangan sensitif dan pendekatan pertahanan dari domain aplikasi seperti pengenalan biometrik ke dalam kerangka teori pembelajaran. Dalam konteks ini, kita juga dapat mengeksplorasi teknologi peningkatan privasi yang menerapkan kontrol pada berbagi data dan analisis kolaboratif dalam pengukuran Internet. Di sini, kita

dapat merancang analisis data, penemuan pengetahuan, dan algoritma pembelajaran mesin untuk kerangka kerja berbagi data. Di dalamnya, batasan domain dapat dimodelkan sebagai fungsi biaya yang merugikan, dan batasan desain dapat dimodelkan sebagai fungsi hasil yang merugikan.

Informasi keamanan dapat direpresentasikan dengan jaringan yang kompleks. Kemudian garis dasar pembelajaran mendalam dapat berfungsi dalam hal proses penambangan data dan fitur pembelajaran mesin. Pembelajaran gabungan yang aman dan dapat diskalakan juga dapat diterapkan ke dalam sistem terdistribusi dan sistem basis data. Penelitian kami dalam pembelajaran adversarial menyediakan kerangka kerja untuk menganalisis keamanan dan privasi dalam pembelajaran mesin. Dalam infrastruktur komputasi berkinerja tinggi, kita dapat mengimplementasikannya dalam alat dan kerangka kerja untuk algoritma serial, algoritma paralel, dan komputasi data besar yang terdistribusi.

Strategi Pelatihan Adversarial

Zhou dkk. mensurvei pendekatan teoritis permainan terhadap pembelajaran mesin permusuhan dengan fokus khusus pada aplikasi keamanan siber karena serangan dari musuh aktif. Aplikasi tersebut mencakup deteksi intrusi, deteksi penipuan perbankan, pemfilteran spam, dan deteksi malware. Permainan keamanan yang sesuai dibentuk antara sistem pembelajaran dan musuh yang cerdas di mana kedua pemain berusaha memainkan strategi respons terbaik yang memaksimalkan keuntungan mereka. Setiap pemain menentukan strategi optimal berdasarkan prediksi pilihan strategi lawan.

Proses pembelajaran permusuhan kemudian menghasilkan pengklasifikasi yang kuat sehingga tidak terlalu rentan untuk disesatkan oleh manipulasi permusuhan. Permainan simultan dan permainan berurutan adalah permainan keamanan paling populer untuk mempelajari interaksi strategis antara musuh cerdas dan sistem pembelajaran. Dalam permainan Stackelberg, lawan atau pembelajar adalah pemimpin permainan. Mereka dapat diperluas ke permainan pemimpin tunggal dengan banyak pengikut seperti permainan Bayesian Stackelberg.

Alpcan dkk. membahas pengambilan keputusan dalam ruang fitur berdimensi tinggi bervolume tinggi dengan permainan strategis dalam sistem yang kompleks seperti jaringan komunikasi, jaringan listrik pintar, sistem cyber-fisik, dan keamanan jaringan. Di sini, pembelajaran permusuhan teoritis permainan dapat digunakan dalam optimasi non-linier dalam sistem variabel yang besar di mana pemain memiliki informasi dan sumber daya yang terbatas untuk secara efektif mengidentifikasi parameter preferensi mereka dengan biaya rendah dalam pengaturan online yang dinamis. Tantangan atas keterbatasan informasi tersebut semakin meningkat dalam kumpulan data berdimensi tinggi untuk analisis data besar.

Pada saat yang sama, masalah optimasi non-linier menghadirkan tantangan komputasi. Jadi rumusan teori permainan yang diusulkan berfokus pada representasi data berdimensi rendah yang direduksi dengan transformasi linier seperti proyeksi acak dan metode pengambilan sampel dalam sampai pada konsep solusi keseimbangan Nash. Proyeksi acak baru dibuat untuk kira-kira mempertahankan jarak antar titik dan hasil kali dalam yang

digunakan dalam algoritma pembelajaran. Permainan strategis besar dan permainan kuadrat dirancang untuk mendekati konsep solusi dalam optimasi.

Oleh karena itu, permainan strategis berskala besar dapat mempelajari inferensi statistik dalam pembelajaran mesin dalam konteks permusuhan. Dalam keamanan komputer, pendekatan terhadap pembelajaran yang aman dapat dibandingkan dengan kerangka teoritis untuk komputasi multi-pihak yang aman dan privasi diferensial. Mereka juga dapat membuat konsep beberapa strategi serangan berulang yang berupaya menyesatkan pengklasifikasi yang digunakan dalam deteksi spam, pendeteksi worm polimorfik, dan pendeteksi anomali jaringan.

Penyesalan yang meminimalkan pembelajar dan metode statistik non-parametrik yang dapat menangani sejumlah besar parameter juga sebanding dengan algoritma pembelajaran adversarial. Jadi desain algoritma pembelajaran adversarial teoritis permainan dan analisis permainan pembelajaran mesin adversarial dapat mengambil manfaat dari teori optimasi yang ditemukan dalam statistik yang kuat, teori pembelajaran online, dan teori minimalisasi penyesalan.

Li dkk. memperluas pembelajaran mesin permusuhan untuk memperhitungkan kendala operasional dalam keputusan acak. Pemisahan konseptual dibuat antara mempelajari preferensi penyerang dan keputusan operasional sehubungan dengan prediksi pembelajaran mesin. Tugas klasifikasi permusuhan dengan penguatan dipisahkan menjadi tugas pembelajaran untuk memprediksi preferensi serangan dan tugas mengoptimalkan kebijakan operasional yang secara eksplisit mematuhi batasan operasional pada prediktor.

Kemudian strategi respons terbaik pihak lawan dihitung sebagai keputusan operasional yang diacak. Data pelatihan ditafsirkan sebagai preferensi penyerang yang terungkap dalam penghindaran permusuhan. Representasi dasar diusulkan untuk memperkirakan secara kompak fungsi keputusan operasional dalam program linier. Kebijakan operasional yang diacak secara eksplisit mematuhi batasan operasional. Pendekatan pembangkitan kendala yang berulang-ulang menciptakan respons terbaik bagi pihak lawan.

Oleh karena itu, cara berprinsip untuk memasukkan pengacakan ke dalam klasifikasi adversarial dengan teknik pembelajaran mesin siap pakai diperkenalkan. Program linier dasar yang memperkirakan keputusan operasional memiliki jumlah variabel dan batasan yang eksponensial. Perkiraan skalabelnya diperoleh dari representasi fungsi Boolean Fourier yang dikombinasikan dengan pembangkitan batasan untuk menghitung keputusan operasional acak berdasarkan batasan anggaran. Hart dkk. membahas strategi adaptif dalam permainan berulang yang mencakup permainan fiktif yang mulus dan pencocokan penyesalan. Distribusi permainan empiris adalah keseimbangan yang berkorelasi dengan hasil vektor. Permainan seperti ini dapat digunakan untuk pembelajaran diskriminatif adaptif dengan optimalisasi berkelanjutan dalam permainan pembelajaran adversarial teoritis.

Jia dkk. menggabungkan Internet of Things (IoT) dan analisis data tingkat lanjut untuk mempelajari fungsi utilitas agen teoritis permainan dari data. Metode berbasis data tersebut memperoleh model utilitas dari keputusan yang diamati dalam keseimbangan dengan algoritma komputasi untuk inferensi statistik seperti optimasi invers dan kontrol optimal

invers. Mereka dapat menggabungkan kerangka kerja tangkas untuk pengembangan analitik guna memprediksi perilaku agen dan merancang insentif untuk mencapai tujuan teoritis permainan dalam perkiraan untuk agen dalam permainan sumber daya bersama.

Perilaku agen teoritis permainan yang dimodelkan dalam permainan energi dapat diterapkan pada program respons permintaan di pasar seperti yang dibentuk oleh sistem komoditas, energi, dan transportasi online. Misalnya, dalam permainan energi gedung pintar, penghuninya mengonsumsi sumber daya bersama seperti penerangan, pemanas, ventilasi, dan AC. Permainan energi kemudian memberikan insentif untuk menggunakan energi secara efisien melalui imbalan uang atas konsumsi energi. Fungsi utilitas untuk masing-masing agen diperkirakan dalam permainan non-kooperatif berdasarkan tindakan historis mereka.

Pendakian gradien yang diproyeksikan (PGA) digunakan untuk mengembangkan strategi serangan keracunan yang optimal dalam algoritma adversarial pembelajaran utilitas. Algoritme permusuhan pembelajaran utilitas mensintesis titik serangan berbahaya untuk meniru perilaku normal. Kondisi stasioner pada perilaku normal dinyatakan sebagai residu regresi. Keputusan dan perilaku agen kemudian dihitung pada ekuilibrium Nash dari estimasi fungsi utilitas. Model ancaman yang kuat dirancang berdasarkan prinsip Kerckhoffs yang kemudian ditargetkan oleh musuh yang meracuni kumpulan data pelatihan untuk menyesatkan prediksi dan mencapai tujuan jahat.

Gradien dalam PGA ditafsirkan sebagai sensitivitas pelatihan yang menangkap perubahan dalam model utilitas yang dipelajari sehubungan dengan data permusuhan. Demikian pula, sensitivitas pengujian memperkirakan variasi dalam keseimbangan Nash sehubungan dengan parameter fungsi utilitas suatu agen. PGA kemudian terbukti mengurangi kekuatan prediksi fungsi utilitas yang dipelajari. Eksperimen dilakukan pada kumpulan data game sintetis dan dunia nyata. Penyerang yang kuat kemudian dapat menghasilkan serangan keracunan berbahaya yang mengakibatkan kesalahan besar dalam perilaku dan prediksi agen yang tidak dapat dibedakan dari tindakan normal. Terdapat trade-off antara kemanjuran serangan dan kemampuan deteksi. Pertukaran ini digunakan untuk mengusulkan mekanisme pertahanan dalam mendeteksi tindakan jahat dan menerapkan metode pembelajaran yang kuat untuk estimasi utilitas.

Dritsoula dkk. mengeksplorasi permainan klasifikasi penyusup sebagai permainan keamanan di mana pembela strategis mengklasifikasikan penyusup sebagai mata-mata atau spammer berdasarkan solusi permainan non-zero sum. Pengklasifikasi dirancang untuk mendeteksi serangan server file dan server email. Fungsi objektif pembela HAM mengacak serangkaian ambang batas untuk menyeimbangkan deteksi yang terlewat dan alarm palsu. Penyerang melakukan trade-off antara meningkatkan kekuatan serangan dan peluang tertangkap. Ekuilibria Nash dalam strategi campuran dihitung untuk permainan jumlah bukan nol dalam waktu polinomial.

Fokus kerangka teoritis permainan ini adalah klasifikasi penyerang, bukan deteksi intrusi. Spammer adalah pemain non-strategis yang diwakili dengan distribusi probabilitas yang tetap dan diketahui. Mata-mata adalah pemain strategis yang memilih jumlah serangan terhadap target utama berdasarkan fungsi biaya permusuhan. Permainan keamanan

dibandingkan dengan permainan sinyal dan dinamis dengan beberapa tahapan di mana para pemain memperbarui keyakinan dan distribusi mereka berdasarkan statistik Bayesian. Matriks biaya mata-mata mempengaruhi strategi respons terbaik pembela HAM dalam formulasi imbalannya. Distribusi serangan spammer mempengaruhi strategi keseimbangan Nash mata-mata tersebut. Strategi ekuilibrium Nash dari pemain bertahan adalah pengacakan pada serangkaian ambang batas yang berdekatan di mana parameter permainan memenuhi kondisi untuk pengacakan di antara rangkaian ambang batas tersebut.

4.2 TEORI PERMAINAN MANIPULASI PEMBELAJARAN MESIN

Dalvi dkk. menganalisis kinerja pengklasifikasi dengan melihat klasifikasi sebagai permainan dengan pengklasifikasi beradaptasi dengan musuh, yang bertujuan untuk membuat pengklasifikasi menghasilkan negatif palsu. Di sini, musuh yang sensitif terhadap biaya dikombinasikan dengan pengklasifikasi yang sensitif terhadap biaya untuk menentukan klasifikasi permusuhan dalam kerangka teori permainan. Dalam klasifikasi adversarial, proses menghasilkan data diperbolehkan berubah seiring waktu sehingga perubahan data dapat dinyatakan sebagai fungsi parameter pengklasifikasi.

Pengklasifikasi kemudian diasumsikan memaksimalkan hasil yang diharapkan atas parameter biaya musuh. Pada gilirannya, strategi musuh adalah menemukan perubahan fitur klasifikasi yang memaksimalkan keuntungan yang diharapkan musuh. Ekuilibria Nash didemonstrasikan untuk permainan berurutan dan permainan berulang dimana parameter kedua pemain diketahui satu sama lain.

Lowd dkk. memperkenalkan algoritma permusuhan untuk mempelajari batas keputusan pengklasifikasi linier. Kerangka pembelajaran ACRE digunakan untuk menentukan apakah musuh dapat belajar cukup efisien dalam mengalahkan pengklasifikasi dengan meminimalkan fungsi biaya permusuhan linier. Biggio dkk. mendefinisikan serangan keracunan terhadap mesin vektor dukungan (SVM) dengan memasukkan contoh permusuhan ke dalam data pelatihan. Prosedur pendakian gradien menghitung contoh permusuhan sebagai maksimum lokal dari permukaan kesalahan non-cembung SVM.

Bruckner dkk. mengusulkan permainan prediksi untuk memodelkan interaksi antara pembelajar yang membangun model prediktif dan generator data yang mengendalikan proses pembuatan data. Kantarcioglu dkk. merancang keseimbangan Nash subgame yang sempurna, yang mengoptimalkan pemilihan atribut dengan fungsi biaya dalam permainan Stackelberg klasifikasi adversarial. Liu dkk. memodelkan perilaku bersaing antara musuh rasional dan penambang data kotak hitam sebagai permainan Stackelberg yang berurutan. Chivukula dkk. menyempurnakan proposal untuk model pembelajaran mendalam, sementara Yin et al. memperluasnya untuk skenario serangan yang jarang. Zhou dkk. mengeksplorasi kerangka permainan bersarang, di mana strategi permusuhan dipilih berdasarkan probabilitas membuat prediksi tentang batas keputusan pengklasifikasi dalam permainan pemimpin tunggal-banyak pengikut.

Kantarcioglu dkk. mengembangkan kerangka teori permainan untuk menganalisis aplikasi pembelajaran permusuhan seperti deteksi intrusi dan deteksi penipuan. Di dalamnya,

kinerja keseimbangan pengklasifikasi menunjukkan keberhasilan atau kegagalannya. Permainan keamanan diselesaikan untuk memprediksi keadaan akhir keseimbangan. Kondisi di mana keseimbangan ada memperkirakan kinerja pengklasifikasi dan perilaku musuh. Jadi keseimbangan teoritis permainan memberikan panduan dalam membangun pengklasifikasi yang berguna dalam penambangan data dan penemuan pengetahuan.

Konsep solusi permainan keamanan diselesaikan dengan simulasi anil stokastik dan integrasi Monte Carlo untuk menemukan strategi keseimbangan. Dalam evaluasi eksperimental, biaya klasifikasi untuk kesalahan klasifikasi label positif ditemukan jauh lebih tinggi dibandingkan biaya klasifikasi untuk kesalahan klasifikasi label negatif. Dengan biaya kesalahan klasifikasi yang sama dan ukuran populasi yang sama untuk label positif dan negatif, pengklasifikasi akan meminimalkan jumlah kesalahan kesalahan klasifikasi.

Liu dkk. memodelkan interaksi dan hasil antara musuh yang cerdas dan sistem pembelajaran sebagai permainan Stackelberg linier non-kooperatif berurutan dua orang. Konsep solusi keseimbangan Nash diperoleh dengan menyelesaikan masalah optimasi diskrit dan kontinu dalam permainan Stackelberg yang dirumuskan sebagai masalah pemrograman bilevel. Ruang strategi bagi musuh bisa terbatas dan tidak terbatas. Dalam kasus yang tidak terbatas, para pemain dalam game tidak perlu mengetahui semua fungsi pembayaran. Algoritme genetika memecahkan permainan Stackelberg untuk kasus tak terbatas.

Prediksi tak terduga dalam distribusi data pelatihan dapat dipelajari dalam kaitannya dengan pembelajaran adversarial di mana gangguan dalam data pelatihan dapat mengubah cara jaringan dalam memprediksi dengan cara yang tidak diinginkan. Rakhlin dkk. memberikan pembelajaran online algoritma tanpa penyesalan dengan teori permainan. Ini disebut penurunan cermin optimis. Ini digunakan untuk masalah optimasi online dalam urutan yang dapat diprediksi. Ia menyatu ke kesetimbangan minimax dalam permainan matriks jumlah nol yang terbatas dalam waktu logaritmik. Jadi jaminan penyesalan kotak hitam dapat ditemukan untuk analisis prediktif pada urutan sewenang-wenang yang mewakili kinerja kasus terburuk dari interaksi teoritis permainan dalam pembelajaran mendalam yang bermusuhan.

Mereka dapat digunakan untuk merancang fungsi kerugian khusus untuk masalah pengambilan sampel, prediksi, dan pengoptimalan dalam pembelajaran mendalam adversarial. Urutan yang tidak berbahaya ini disebabkan oleh kelancaran pengoptimalan bagian dalam dan struktur pengoptimalan titik pelana yang sesuai dalam formulasi game. Di sini, urutan yang dapat diprediksi/normal mengacu pada label kelas yang tidak berbahaya, dan urutan yang tidak terduga/tidak normal mengacu pada label kelas yang berlawanan. Prediktabilitas barisan berasal dari prediktabilitas gradien dalam kriteria optimasi dan konvergensi untuk menemukan nilai minimax.

Tujuan pengoptimalannya adalah untuk meminimalkan penyesalan kumulatif antara urutan yang dapat diprediksi dan yang berlawanan yang disisipkan di antara urutan hasil teoritis permainan untuk semua pemain yang berpartisipasi dalam permainan keamanan. Dalam algoritme penurunan cermin optimis, divergensi Bregman yang ditentukan pada fungsi kerugian khusus menentukan ukuran langkah dalam rangkaian prediktif dan menyesuaikan

ukuran langkah dengan rangkaian yang diamati sejauh ini. Kami juga dapat menggunakan fungsi kerugian khusus pada momen dan kumulan distribusi data untuk mengeksplorasi nilai prediksi “abnormal” dan “penjelasan” fitur sebab akibat yang sesuai untuk kecerdasan buatan yang dapat dijelaskan.

Di sini, kita dapat bereksperimen dengan prosedur pemilihan fitur dan desain fungsi kerugian yang melibatkan varian umum, jaringan Bayesian non-parametrik, dan struktur sebab akibat probabilistik dalam regresi/interpolasi jaringan dalam multivariat antar urutan. Di sini, mekanisme berbasis perhatian dan generatif dalam pembelajaran mendalam dapat menjadi dasar regresi untuk analisis varians dan evaluasi model dalam prediksi terstruktur multivariat. Kami juga dapat menggunakan statistik deskriptif, statistik ringkasan, statistik memadai, dan statistik urutan yang sesuai untuk analisis varians dalam urutan yang dapat diprediksi. Motwani dkk. adalah teks standar tentang algoritma acak untuk masalah pembelajaran online dalam pembelajaran adversarial teoritis permainan.

Blum dkk. menyajikan permainan berbagi sumber daya secara berurutan untuk mencapai kesejahteraan sosial dengan informasi yang diumumkan secara publik dan menjaga privasi. Kasus penggunaan yang memotivasi diambil dari pengaturan multi-agen dalam pengambilan keputusan keuangan di mana pemain memainkan permainan dengan informasi yang tidak sempurna. Ide kesejahteraan sosial dalam game ini bergantung pada tindakan para pemain di masa lalu. Penerapannya kemudian ditampilkan pada permainan penjadwalan mesin dan pembagian biaya. Penyebaran informasi pribadi yang terhormat adalah mekanisme pertahanan permusuhan yang direkomendasikan dan merupakan titik temu antara desain mekanisme dan pembelajaran mesin yang menjaga privasi.

Mekanisme perlindungan privasi mengumpulkan informasi dari pemain untuk menghitung perkiraan keseimbangan berkorelasi yang memberikan saran kepada pemain tentang permainan optimal sesuai dengan jenis dan perilaku pemain. Pertukaran antara peningkatan informasi perkiraan swasta tentang keadaan permainan dan penurunan kesejahteraan sosial dianalisis dengan dinamika respons terbaik dalam permainan pencocokan serakah yang memiliki fungsi biaya yang berisik. Tidak seperti pembelajaran mendalam yang bersifat adversarial, fungsi biaya yang berisik diperkirakan berdasarkan keadaan permainan dalam game, bukan hasil dari permainan yang mengabstraksi nilai dari tindakan pemain. Dengan demikian, kita dapat menambah fungsi pembayaran pemain dengan persamaan keadaan sistem dinamis untuk menghasilkan kontrol stokastik dalam interaksi teoretis permainan dari pembelajaran mendalam yang bermusuhan.

Dalam penelitian kami, kami mengevaluasi algoritma pembelajaran adversarial multi-label dalam pengaturan optimasi stokastik. Kami secara empiris menghasilkan manipulasi permusuhan pada keseimbangan Nash dalam permainan Stackelberg sekuensial jumlah konstan dan jumlah variabel. Ruang strategi musuh kita ditentukan oleh parameter evolusi dan parameter variasi yang dipelajari dari distribusi data masukan. Oleh karena itu, manipulasi adversarial optimal yang ditemukan oleh musuh kita menentukan model generatif untuk data adversarial yang ditemukan di ruang data masukan pengklasifikasi.

Selanjutnya, kami mendefinisikan fungsi biaya adversarial pada ruang strategi yang mengkode distribusi data asli. Fungsi pembayaran permusuhan kami dioptimalkan oleh algoritma simulasi anil yang mengacak perubahan langkah dalam ruang strategi musuh. Pengacakan dalam strategi serangan permusuhan kami juga ditentukan oleh ruang laten yang merekonstruksi distribusi data asli dengan autoencoder variasional (VAE). Fungsi biaya adversarial yang diusulkan (variasi non-linier non-cembung) mengarah pada regularisasi yang lebih baik dari fungsi hasil adversarial yang menyatu ke ekuilibrium Nash dalam permainan Stackelberg kami.

Optimasi Bertingkat dalam Pembelajaran mesin Game Theoretical

Berdasarkan interpretasi visual pengklasifikasi probabilistik, Di Nunzio et al. mensurvei gamifikasi teknik pembelajaran mesin yang diawasi untuk memberi label objek dengan biaya terjangkau dan tidak memakan waktu. Model penetapan harga dirancang untuk membangun pengklasifikasi yang cukup akurat dengan sampel objek berlabel berukuran kecil yang kinerjanya sebanding dengan algoritma klasifikasi canggih. Permainan ini disusun menjadi sepuluh level sesuai dengan kriteria keterpisahan antara kelas positif dan kelas negatif. Sasaran di setiap level adalah menemukan pengklasifikasi terbaik yang memaksimalkan skor F1 dengan sumber daya komputasi paling sedikit.

Liaghati dkk. membahas pendekatan optimasi maksimal untuk desain sistem tangguh yang beroperasi dalam lingkungan ko-evolusi. Di sini, manipulasi permusuhan dianggap sebagai perilaku muncul tak terduga yang ditunjukkan oleh sistem yang kompleks. Perilaku yang muncul muncul karena adanya keragaman, konektivitas, interaktivitas, dan adaptasi suatu sistem di lingkungannya. Pemrograman matematika digunakan dalam desain sistem untuk mengoptimalkan sistem sehubungan dengan kompleksitas dan pengorbanan dalam lingkungan operasi. Kompleksitas berbagai interaksi, perumusan masalah, pemangku kepentingan, dan lingkungan operasional yang berbeda. Kompleksitas tersebut mencakup banyak pemain, interaksi yang berbeda antar pemain, perumusan masalah pembelajaran, pemangku kepentingan penambangan data, dan lingkungan operasional untuk pembelajaran mesin.

Dalam konteks ini, contoh-contoh permusuhan mengarah pada peristiwa angsa hitam (black swan) dalam sistem yang kompleks. Ketahanan sistem dan ketahanan terhadap persaingan kemudian dicapai melalui desain yang bijaksana dan terinformasi yang menjadikan sistem efektif dan efisien dalam berbagai konteks. Definisi ketahanan diperluas untuk mempertahankan kemampuan dalam menghadapi gangguan dengan menyerap tekanan eksternal. Ketahanan didefinisikan dalam kaitannya dengan kemungkinan pemulihan berdasarkan prognosis yang benar dan diagnosis yang benar. Pemrograman matematika untuk desain sistem mencakup pemrograman non-linier bilangan bulat campuran, algoritma evolusioner dengan banyak batasan dan fungsi kebugaran non-linier, optimasi Bayesian untuk persyaratan keamanan yang tidak pasti, dan optimasi kuat minimax/maximin yang cocok untuk masalah non-linier.

Penerapan teori permainan pada rekayasa sistem mengarah pada perancangan sistem di mana lingkungan operasional yang berlawanan tidak dapat dikendalikan dan tidak pasti.

Lingkungan operasi dalam desain sistem tangguh diuji terhadap distribusi seragam yang diperoleh berdasarkan prinsip entropi maksimum. Sistem pertahanan udara dipilih sebagai representasi dari sistem kompleks yang beroperasi di lingkungan yang bermusuhan dengan biaya komputasi yang terkait. Yair dkk. menafsirkan prosedur statistik yang disebut pembelajaran divergensi kontrastif (CD) sebagai pembelajaran permusuhan di mana diskriminator mengklasifikasikan apakah rantai Markov yang dihasilkan dari model statistik telah dibalik waktu.

CD juga dikontraskan dengan pembelajaran dalam jaringan permusuhan generatif (GAN) untuk menyimpulkan bahwa aturan pembaruan CD tidak dapat dinyatakan dalam gradien fungsi tujuan tetap apa pun. CD memiliki keunggulan empiris dalam estimasi kemungkinan maksimum (MLE) karena rantai Markov pendek yang diinisialisasi pada sampel data yang ditemukan di berbagai domain aplikasi. CD menyesuaikan distribusi kontras untuk menghasilkan sampel yang dekat dengan manifold namun menempuh jarak yang jauh di sepanjang manifold.

Pengoptimalan bertingkat cocok untuk permainan kompetitif di mana tidak ada pemain berpeluang. Pengoptimalan multi-tahap cocok untuk permainan kooperatif di mana semua pemain menerima hasil yang sama, namun ada pemain yang berpeluang. Gerakan para pemain dalam permainan kooperatif bergantian antara pemain yang bekerja sama dan pemain yang berpeluang. Masalah pengambilan keputusan dalam pembelajaran adversarial teori permainan dapat dirumuskan sebagai masalah optimasi dalam teori optimasi bertingkat dan multi-tahap yang mencakup banyak pengambil keputusan independen, proses pengambilan keputusan berurutan atau multi-tahap, dan beberapa tujuan yang mungkin saling bertentangan.

Di sini, optimasi bertingkat memiliki banyak tahapan, banyak tujuan, dan banyak pengambil keputusan. Sebaliknya, optimasi bertingkat menggeneralisasi pemrograman matematika dengan model masalah keputusan dan kelas kompleksitas untuk menentukan pergerakan setiap pemain sebagai solusi masalah optimasi dalam pembelajaran adversarial teoritis permainan. Aplikasi praktis untuk optimasi multi-tahap tersebut mencakup sistem keputusan hierarki seperti lembaga pemerintah dan perusahaan besar dengan banyak anak perusahaan, sistem optimasi terkontrol seperti jaringan listrik, dan sistem biologis. Teori optimasi multilevel juga memungkinkan kita untuk mengukur trade-off sumber daya komputasi antara privasi dan keamanan, biaya musuh dan biaya pembelajar, dan skenario serangan dan mekanisme pertahanan dalam pembelajaran permusuhan teoritis dalam hal formulasi dualitas dalam optimasi seperti pemisahan versus optimasi, invers.

Optimasi versus optimasi ke depan, harga versus sensitivitas, dan fungsi primal versus ganda dalam optimasi. Chalkiadakis dkk. menulis buku tentang teori permainan kooperatif tentang perilaku strategis agen yang berkepentingan sendiri dengan kesepakatan yang mengikat di antara mereka. Aspek komputasi teori permainan kooperatif dirangkum seperti utilitas yang dapat ditransfer dalam permainan, konsep solusi seperti nilai Shapley, representasi kompak untuk permainan, dan konsep solusi komputasi yang efisien untuk permainan. Algoritme teoritis permainan yang cocok untuk pembelajaran permusuhan

mencakup struktur koalisi yang memaksimalkan kesejahteraan, metode untuk membentuk koalisi dalam kondisi ketidakpastian, dan algoritma tawar-menawar.

Sebuah permainan dapat diartikan sebagai model hubungan multi-agen antara tindakan agen dan insentif. Ketika agen mementingkan diri sendiri, permainan memodelkan proses optimasi untuk menggambarkan proses yang tidak pasti dengan model probabilistik yang mendasarinya. Di sini, optimasi multi-tujuan secara bersamaan mengoptimalkan beberapa tujuan untuk menemukan kurva Pareto-optimal yang merupakan sekumpulan titik di mana setiap fungsi tujuan tidak dapat bertambah besar dengan menurunnya fungsi tujuan lainnya. Dalam teori permainan, setiap fungsi tujuan dimiliki oleh agen yang terpisah, dan variabel keputusan dipartisi ke dalam domain fungsi tujuan masing-masing pemain.

Di sini, permainan pengejaran dan penghindaran serta permainan penyebaran sumber daya strategis dapat dirumuskan untuk peperangan algoritmik. Permainan ekonomi berhubungan dengan lelang; pembelian/penjualan dapat digunakan dalam periklanan komputasi, pengadaan sumber daya, analisis pasar saham, dan penemuan harga dinamis. Permainan grafis dirancang berdasarkan pembentukan jaringan antar pemain di jaringan sosial, korporat, dan P2P. Permainan rekreasional seperti catur, catur, go melibatkan informasi lengkap tentang lawannya. Deb dkk. membahas aplikasi dunia nyata dari optimasi multi-tujuan evolusioner (EMO). Solusi pareto-optimal ditemukan dengan metodologi optimasi EMO yang dapat menangani sejumlah besar tujuan, biaya komputasi yang besar, dan kesulitan dalam visualisasi ruang tujuan.

Prosedur EMO berkembang ke wilayah Pareto-optimal dengan secara adaptif menemukan interaksi dimensi rendah yang benar. Hal ini diimplementasikan dengan algoritma pengurutan GA atau NSGA-II yang elitis dan tidak didominasi yang dapat menskalakan banyak tujuan. Pareto optima seperti itu dapat memodelkan ruang strategi acak permainan dalam pembelajaran persaingan teoritis permainan. Kita dapat merumuskan manipulasi permusuhan sebagai optimasi stokastik dan parameter pengambilan sampel acak dari autoencoder variasional yang menghasilkan data permusuhan dalam permainan Stackelberg. Permainan semacam ini dirancang untuk menyesatkan musuh dengan statistik yang kuat dalam hal algoritma EMO seperti simulasi anil (SA) dan Alternating Least Squares (ALS). Skalar optima di SA digunakan untuk menghasilkan vektor optima di ALS. Kekuatan dan relevansi skenario serangan kami ditentukan oleh performa model pembelajaran mendalam yang diserang.

Zhang dkk. mensurvei titik temu antara algoritma komputasi evolusioner (EC) dan teknik pembelajaran mesin. Mereka termasuk algoritma genetika (GA), pemrograman evolusioner (EP), strategi evolusioner (ES), pemrograman genetika (GP), sistem pengklasifikasi pembelajaran (LCS), evolusi diferensial (DE), estimasi algoritma distribusi (EDA), semut. optimasi koloni (ACO), optimasi gerombolan partikel (PSO), dan algoritma memetika (MA). Teknik pembelajaran mesin yang menggunakan algoritma EC meliputi metode statistik, interpolasi dan regresi, analisis clustering (CA), analisis komponen utama (PCA), desain eksperimental ortogonal (OED), pembelajaran berbasis oposisi (OBL), jaringan syaraf tiruan

(ANN) , mendukung mesin vektor (SVM), penalaran berbasis kasus, pembelajaran penguatan, pembelajaran kompetitif, dan jaringan Bayesian.

Taksonomi dihasilkan untuk membandingkan langkah evolusi yang meningkatkan setiap teknik pembelajaran mesin. Pembelajaran permusuhan teoretis permainan dapat menggunakan langkah-langkah evolusi untuk adaptasi parameter, adaptasi operator, pencarian lokal, dan biaya komputasi untuk menghasilkan metode komputasi numerik dalam teori permainan evolusi dengan musuh evolusioner menyusun algoritma dinamis untuk menghasilkan manipulasi permusuhan. Ruang strategi teoritis permainan untuk pengacakan algoritmik dan manipulasi data dalam penelitian kami ditentukan oleh operator stokastik dalam algoritma evolusioner dan jaringan variasional yang menentukan skenario serangan. Algoritme penelusuran evolusioner dari penelitian saya dapat diperluas ke dalam proses pengambilan keputusan Markov dan automata seluler dengan perluasan prosedur pengoptimalan lokal yang memaksimalkan fungsi hasil yang merugikan.

Di sini, jaringan kepadatan campuran dapat mengekspresikan distribusi data bersyarat pada variabel laten dan label kelas dalam data pelatihan dan data permusuhan. Kami juga dapat mengukur perbedaan informasi antara representasi minimal data pelatihan dan penyematan fitur data adversarial dengan fungsi biaya adversarial berbasis pembelajaran metrik yang mendalam. Kami juga dapat menerapkan distribusi sebelumnya pada faktor laten untuk pembuatan data yang koheren dalam pembelajaran yang diawasi.

4.3 TEORI PERMAINAN PEMBELAJARAN MESIN MENDALAM

Dalam Liu dkk., interaksi antara musuh dan penambang data dimodelkan sebagai permainan zero-sum Stackelberg berurutan dua pemain di mana hasil untuk setiap pemain dirancang sebagai fungsi kerugian yang diatur. Pergerakan setiap pemain didasarkan pada pengamatan permainan terakhir lawan. Musuh secara berulang menyerang penambang data dengan strategi terbaik untuk mengubah data pelatihan asli. Penambang data bereaksi dengan membangqun kembali pengklasifikasi berdasarkan pengamatan penambang data terhadap modifikasi musuh pada data pelatihan. Strategi bermain lawan ditentukan secara mandiri oleh lawan. Permainan ini diulang sampai hasil lawan tidak bertambah atau jumlah iterasi maksimum tercapai.

Masalah maximin untuk optimasi diusulkan dalam Liu et al. diselesaikan tanpa membuat asumsi tentang distribusi yang mendasari data pelatihan dan pengujian. Evaluasi empiris algoritma optimasi dilakukan terhadap data spam gambar dan spam teks. Pengaturan fungsi kerugian yang berbeda menghasilkan jenis pengklasifikasi yang berbeda seperti regresi logistik dengan fungsi kerugian linier log dan mesin vektor dukungan dengan fungsi kerugian engsel. Untuk fungsi kerugian yang dipilih, tujuan optimasi dirumuskan sebagai masalah optimasi cembung tak terbatas.

Masalah optimasi diselesaikan dengan metode wilayah kepercayaan yang meminimalkan fungsi tujuan pada lingkungan koordinat kutub yang dibatasi. Pada ekuilibrium Nash, solusi masalah maximin mencapai tingkat negatif palsu tertinggi dan biaya transformasi data terendah secara bersamaan. Hal ini mengarah pada batasan klasifikasi yang kuat pada

waktu pengujian. Vektor bobot yang dihitung pada ekuilibrium Nash juga memberikan fitur yang lebih kuat terhadap manipulasi data yang merugikan.

Liu dkk. mengusulkan perpanjangan ke Liu dkk. di mana permainan satu langkah digunakan untuk mengurangi waktu komputasi algoritma minimax. Metode satu langkah konvergen ke ekuilibrium Nash dengan memanfaatkan dekomposisi nilai tunggal (SVD). SVD memberikan vektor basis ortogonal atau vektor tunggal yang bertindak sebagai "komponen utama" data pelatihan. Dengan demikian, vektor tunggal mencirikan setiap jenis kelas yang ada dalam data pelatihan. Label data uji kemudian diambil oleh Liu et al. menjadi kelas pelatihan yang menghasilkan vektor residu yang lebih kecil.

Algoritme klasifikasi berbasis SVD ini dianggap sebagai keadaan awal model teoretis permainan. Sasaran musuh adalah mengubah instance kelas target (atau kelas positif) menjadi kelas negatif. Tujuan ini dicapai dalam data pengujian dengan menggeser sejumlah kecil kejadian positif dalam data pelatihan sehingga distribusi kelas condong ke arah label kelas positif. Selain itu, fungsi payoff untuk adversarial dirumuskan sebagai selisih vektor tunggal sebelum dan sesudah manipulasi data adversarial. Dengan demikian, musuh yang rasional tidak hanya berupaya meminimalkan jarak antara distribusi kejadian negatif dan transformasi kejadian positif, namun juga meminimalkan transformasi itu sendiri.

Imbalan bagi musuh kemudian kira-kira diselesaikan menggunakan sub-masalah wilayah kepercayaan yang diselesaikan menggunakan perkiraan subruang sambil menghindari perhitungan mahal dari matriks gradien dan matriks hessian dari fungsi kerugian musuh dan fungsi kerugian pengklasifikasi. Dalam memvalidasi algoritme, contoh adversarial dibuat untuk menghasilkan tingkat positif palsu yang tinggi untuk pengklasifikasi pada langkah awal permainan. Di akhir permainan, tingkat positif palsu dari algoritma pembelajaran dikurangi dengan memperhitungkan manipulasi data yang merugikan.

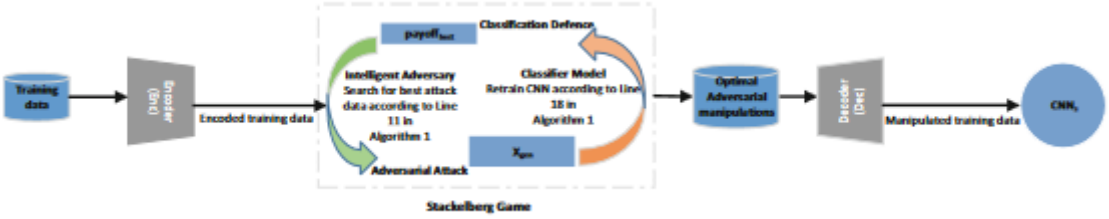
Wang dkk. berasumsi bahwa musuh mengubah fitur pengklasifikasi sesuka hati dan membayar biaya yang sebanding dengan ukuran subset fitur yang telah diubah. Serangan terhadap pengklasifikasi seperti ini disebut serangan fitur renggang di makalah. Masalah optimasi min-maks kemudian dirumuskan sebagai permainan jumlah bukan nol. Dalam permainan non-zero sum, keuntungan pengklasifikasi belum tentu merugikan pihak lawan. Fungsi kerugian yang diregulasi diusulkan untuk penambang data dan musuh untuk menjadikan tujuan permainan sebagai masalah optimasi bilevel cembung.

Regulator l1 dan l2 diperiksa sehubungan dengan model sparse yang diusulkan. Musuh diasumsikan menerapkan perubahan data dengan meminimalkan perubahan fungsi kerugian. Musuh memilih strategi serangan dengan pengetahuan penuh tentang strategi pembobotan fitur penambang data. Dalam mengatur fungsi kerugian musuh, jumlah sampel positif dan jumlah sampel negatif digunakan untuk memperhitungkan ketidakseimbangan dalam data. Kemudian penambang data memilih bobot fitur berdasarkan sampel dalam ruang data yang dimanipulasi. Tujuan dari data miner adalah untuk menentukan batasan keputusan berdasarkan data yang terus dimanipulasi di setiap langkah.

Tujuan musuh adalah menentukan vektor manipulasi berdasarkan anggaran tertentu di setiap langkah. Langkah-langkah ini diulangi secara berurutan hingga konvergensi. Setelah

konvergensi, pengklasifikasi menemukan bobot fitur yang kuat terhadap serangan fitur renggang yang diusulkan. Untuk menyelesaikan tujuan kuadrat terkecil yang diregulasi, manipulasi data pihak lawan diasumsikan dibatasi oleh matriks gangguan sebesar norma l1 yang menetapkan anggaran yang masuk akal (atau akumulasi biaya) sebagai kriteria konvergensi pihak lawan. Elemen matriks perturbasi disesuaikan dengan data masukan melalui validasi silang atas data pelatihan dan pengujian yang diurutkan tepat waktu.

Berbagai permainan tersebut kemudian disimulasikan oleh berbagai pengatur l1 dan l2 pada matriks perturbasi. Pilihan pengatur untuk penambang data menyebabkan trade-off antara ketersebaran dan akurasi serta bias dan varians dari pengklasifikasi. Evaluasi eksperimental memvalidasi bahwa pengklasifikasi teoretis permainan memburuk pada tingkat yang lebih lambat dibandingkan pengklasifikasi biasa pada data masa depan dekat dan masa depan jauh.



Gambar 4.1 Diagram alur yang mengilustrasikan pembelajaran permusuhan teoretis permainan Stackelberg yang bervariasi

Untuk mendapatkan fungsi pembayaran dalam game, kami berasumsi bahwa musuh tidak memiliki pengetahuan tentang lapisan jaringan saraf dalam atau fungsi kerugian dalam model pembelajaran mendalam. Masalah pengoptimalan teoretis permainan yang kami usulkan diselesaikan tanpa membuat asumsi pada distribusi data pelatihan dan pengujian pelajar. Ruang strategi untuk pengacakan algoritmik dan manipulasi data dalam permainan kami ditentukan oleh operator stokastik dalam algoritme evolusioner dan jaringan variasional yang menentukan skenario serangan.

Struktur Keseluruhan Model Pembelajaran dalam Permainan Variasi

Gambar 4.1 adalah diagram alur proses pembelajaran adversarial yang menjelaskan keberadaan musuh variasional dalam pembelajaran yang diawasi. Hasil akhir dari pembelajaran adversarial kami adalah model klasifikasi CNN CNNsecure (selanjutnya disingkat CNNs) yang tahan terhadap serangan adversarial. Kami menghasilkan data permusuhan dalam permainan Stackelberg dua pemain antara musuh dan pengklasifikasi. Musuh membuat model variasi dengan mencari manipulasi permusuhan pada data pelatihan yang dikodekan.

Setiap parameter statistik dari data pelatihan yang dikodekan dicari berdasarkan prosedur simulasi anil (SA). Agregasi manipulasi permusuhan ke semua parameter statistik dalam data pelatihan yang dikodekan dioptimalkan berdasarkan prosedur kuadrat terkecil bergantian (ALS). Pengoptimalan ALS dipanggil setiap kali musuh menghasilkan data musuh

Xgen di game Stackelberg. Xgen bertindak sebagai data validasi untuk pengklasifikasi yang diserang. Untuk setiap Xgen, pengklasifikasi mengoptimalkan ulang bobot pelatihannya untuk memperbarui dirinya sendiri.

Hasil dari interaksi teoritis permainan antara gerakan terbaik pembelajar dan pengklasifikasi diukur dengan hasil terbaik lawan. Musuh melibatkan pengklasifikasi dalam permainan Stackelberg selama hasil terbaiknya meningkat. Penurunan payoffbest menunjukkan bahwa kondisi keluar ekuilibrium Nash telah tercapai pada permainan Stackelberg. Di akhir permainan, musuh memiliki manipulasi adversarial yang optimal dari Xgen terbaru. Manipulasi seperti itu diterapkan pada data pelatihan untuk mendapatkan data pelatihan yang diserang. Kemudian proses pembelajaran pengklasifikasi menambahkan data yang diserang ke dalam data pelatihan asli sehingga CNN dapat dilatih ulang secara optimal oleh serangan musuh kita. Saat pengklasifikasi CNN dilatih di ruang data asli, musuh menghasilkan manipulasi data di ruang data yang disandikan. Representasi variasional dari ruang data yang dikodekan memungkinkan musuh mengusulkan model generatif untuk manipulasi permusuhan.

Perbedaan Metode Kami dan GAN

Meskipun GAN dan metode kami didasarkan pada kerangka teori permainan dan keduanya mencari keseimbangan Nash, keduanya memiliki beberapa perbedaan besar. Pada bagian ini, kami merangkum perbedaan-perbedaan tersebut menjadi tiga aspek, antara lain konstruksi model, optimasi model, dan hasil optimasi. Konstruksi Model Pertama, permainan Stackelberg variasional kami adalah masalah jumlah variabel, sedangkan GAN membuat permainan jumlah konstan. Kedua, metode kami mendefinisikan musuh sebagai pemimpin permainan, sedangkan GAN dipimpin oleh generator. Terakhir, GAN menentukan skenario serangan untuk menemukan model generatif yang mendasari distribusi data tertentu, sementara kami mengoptimalkan fungsi imbalan yang merugikan dengan parameter serangan evolusioner yang mendefinisikan skenario serangan kami dalam ruang strategi acak.

Optimasi Model Pertama, GAN memecahkan masalah optimasi cembung dengan algoritma optimasi berbasis gradien. Sebaliknya, kami memecahkan masalah optimasi stokastik dengan algoritma simulasi anil. Secara khusus, ekuilibrium Nash permainan kami dihitung dengan menyelesaikan masalah optimasi non-cembung. Kedua, selama pelatihan permusuhan teoretis game, kami menanyakan CNN tentang kinerja serangan errorpos(w), sementara GAN menanyakan CNN untuk membedakan antara data “asli” dan data “palsu”. Hasil Pengoptimalan Inilah perbedaan paling signifikan antara GAN dan model kami. Pertama, tujuan GAN pada ekuilibrium Nash adalah untuk mempelajari model generatif yang meniru distribusi data asli, sedangkan metode kami mempelajari manipulasi adversarial optimal ($\alpha\mu$, $\alpha\sigma$) yang bukan merupakan distribusi data asli yang sebenarnya tetapi merupakan manipulasi $*$to distribusi aslinya.

Kedua, diskriminator GAN pada keseimbangan Nash tidak dapat mengklasifikasikan antar label, sementara pengklasifikasi kami kuat terhadap manipulasi permusuhan, dan kinerja pertahanannya diukur dengan errorP os(w). Ketiga, dalam skenario serangan kotak hitam kami, usulan yang berbeda mengenai fungsi imbalan yang merugikan dan fungsi biaya

yang merugikan menyebabkan keseimbangan Nash yang berbeda untuk mempelajari fungsi tujuan dan manipulasi permusuhan yang terkait. Sebaliknya, GAN selalu berusaha menyatu dengan distribusi data pelatihan.

Perbandingan Model Pembelajaran Mesin Mendalam Game Theoretical

Dalam penelitian kami, pembelajaran mendalam adversarial teori permainan diterapkan pada masalah klasifikasi keamanan dunia maya dalam tahap pelatihan dan tahap pengujian. Masalah seperti itu mempelajari manipulasi fitur, biaya kesalahan klasifikasi, dan ketahanan distribusi dalam aplikasi pembelajaran adversarial. Contoh permusuhan kemudian dibuat melalui eksperimen pada fungsi kerugian dalam pembelajaran mendalam. Fungsi kerugian permusuhan dan prosedur pelatihan dalam penelitian kami dapat diterapkan pada studi tentang pembelajaran mendalam yang dapat dipercaya dalam penerapan sistem cyber-fisik.

Mereka dapat mensimulasikan perlindungan, risiko, dan tantangan keamanan dunia maya seperti optimasi komputasi dan masalah inferensi statistik. Untuk analisis sensitivitas pada data besar tersebut, kita dapat menganalisis metrik validasi analitik yang menyesuaikan parameter jaringan neural dalam berdasarkan tren kesalahan klasifikasi dalam prediksi terstruktur. Metrik validasi umum untuk tujuan ini mencakup matriks konfusi, kurva perolehan presisi, kurva ROC, kurva peningkatan, dan statistik kappa. Kita juga dapat menentukan struktur data sinopsis pada tensor dan grafik untuk mendapatkan fitur pembelajaran mesin adversarial. Minat kami adalah pada struktur data yang membantu penelusuran kesamaan dan pembelajaran metrik di seluruh distribusi probabilitas untuk mengevaluasi keamanan algoritme pembelajaran mesin berdasarkan paradigma pembelajaran mesin desain-untuk-keamanan, bukan paradigma desain-untuk-kinerja tradisional.

Dalam konteks ini, kausalitas dan stasioneritas rantai Markov dapat digunakan untuk mendefinisikan prinsip maksimalisasi ekspektasi dan panjang deskripsi minimum untuk inferensi statistik dalam distribusi data yang berlawanan. Kami telah menerapkan fitur-fitur yang dipelajari dalam pengelompokan, klasifikasi, dan analisis asosiasi. Mereka dapat diperluas ke pembelajaran fitur untuk prediksi terstruktur, deteksi perubahan, penambangan peristiwa, dan penambangan pola. Di sini, fitur yang dipelajari dapat berupa salah satu fitur sampel, fitur yang dibangun, fitur yang diekstraksi, fitur yang disimpulkan, dan fitur prediktif. Pengoptimalan biaya adversarial pada berbagai jenis fitur yang dipelajari dapat diparameterisasi secara mudah dengan fungsi kerugian khusus dalam model pembelajaran mendalam yang diawasi dan fungsi biaya adversarial dalam pengoptimalan yang kuat.

Matahari dkk. membahas pemilihan fitur dalam pembelajaran mesin dan pengenalan pola yang didasarkan pada fitur teoretis informasi untuk menghilangkan fitur yang berlebihan dan tidak relevan dari data berdimensi tinggi. Kerangka teori permainan kooperatif diusulkan untuk evaluasi fitur dan pembobotan guna mengoptimalkan kinerja pembelajaran pengurangan dimensi. Nilai Shapley mengevaluasi bobot setiap fitur dalam subset fitur yang saling bergantung untuk menghasilkan algoritma pemilihan fitur. Dengan demikian, teori

permainan dapat digunakan untuk analisis relevansi fitur yang dipelajari, saling ketergantungan, dan analisis redundansi dalam pembelajaran mendalam adversarial.

Secara umum, pemilihan fitur teoretis permainan memiliki aplikasi kecerdasan buatan yang memerlukan kombinasi optimal algoritma pemilihan fitur dan pengklasifikasi untuk pemilihan model yang efisien dalam teori pembelajaran mesin yang menangani informasi yang tidak pasti. Matahari dkk. mengembangkan kerangka teori permainan kooperatif untuk mengevaluasi pentingnya fitur relatif. Pengukuran teori informasi diusulkan untuk membedakan antara redundansi, saling ketergantungan, dan independensi antara fitur yang dipelajari. Indeks kekuatan Banzhaf diperoleh untuk setiap fitur. Ini dirata-ratakan pada koalisi subkumpulan fitur yang dimilikinya untuk mengusulkan konteks pemilihan fitur dalam permainan pemungutan suara. Dampak suatu fitur dihitung berdasarkan koalisi pemenang yang mencerminkan relevansi fitur tersebut dengan kelas target.

Gore dkk. menerapkan teori permainan kooperatif pada pemilihan fitur dalam algoritma Relief. Dengan mengasumsikan distribusi probabilitas pada data pelatihan, ukuran teori informasi diperoleh untuk menentukan peringkat fitur yang dipelajari, memilih subset fitur, dan mengukur kontribusi individual dari setiap fitur yang termasuk dalam subset fitur. Pemilihan fitur berdasarkan teori permainan berguna untuk mempelajari konsep target guna mengoptimalkan fungsi kriteria adversarial dan meningkatkan kinerja ketahanan adversarial untuk memperkirakan distribusi kondisi kelas yang mendasari dalam pembelajaran mendalam adversarial.

Dengan menangani kutukan dimensi, hal ini dapat meningkatkan kinerja prediksi, mengurangi persyaratan pengukuran dan penyimpanan, mengurangi kompleksitas pelatihan dan waktu prediksi, dan memberikan pemahaman yang lebih baik tentang proses stokastik yang menghasilkan data permusuhan. Algoritma Relief bertindak sebagai pembeda antara kelas-kelas berbeda yang mengelompokkan subset fitur. Nilai-nilai Shapley menentukan pentingnya fitur di seluruh iterasi koalisi teoritis permainan komputasi algoritme pemilihan fitur. Cohen dkk. merancang algoritma pemilihan kontribusi (CSA) untuk pemilihan fitur berdasarkan kerangka analisis Shapley multiperturbasi (MSA).

Pembelajaran teoretis permainan menilai kegunaan fitur yang dipilih berdasarkan algoritma pencarian seleksi maju atau eliminasi mundur. Algoritme CSA dapat dioptimalkan berdasarkan ukuran kinerja seperti akurasi, tingkat kesalahan yang seimbang, dan area di bawah kurva karakteristik pengoperasian penerima. Permainan koalisi di CSA mengarah pada iterasi teoretis permainan yang memperkirakan nilai Shapley untuk fitur sesuai dengan kerangka MSA. Dalam eksperimen, subset fitur diproduksi untuk menghasilkan pengklasifikasi berperforma tinggi secara empiris.

Kami merancang permainan Stackelberg antara dua pemain di mana salah satu dari dua pemain, seorang penambang data pengikut (pelajar), bertindak sebagai respons terhadap gerakan pemain lain, seorang pemimpin musuh yang cerdas (adversary), dengan tujuan untuk berkumpul pada keadaan keseimbangan yang disebut Keseimbangan Nash. Kami telah merumuskan fungsi tujuan dalam permainan Stackelberg (zero-sum) (berurutan dua pemain) sebagai masalah optimasi bilevel, di mana tujuan teoritis permainan didefinisikan untuk

secara bersamaan mengoptimalkan dua fungsi pembayaran yang mempengaruhi satu sama lain sesuai dengan interaksi pemimpin-pengikut yang memungkinkan pelajar untuk berlatih kembali setelah setiap serangan.

Algoritma pencarian stokastik kemudian dirancang untuk memecahkan masalah optimasi. Kebijakan serangan optimal dirumuskan dalam bentuk operator optimasi stokastik dan algoritma komputasi evolusioner. Fungsi pembayaran musuh menyimulasikan proses serangan mereka, dan pembelajar menyimulasikan proses pembelajarannya. Solusi terhadap proses serangan menentukan kebijakan serangan optimal musuh dalam batasan. Solusi untuk proses pembelajaran menentukan keuntungan pembelajar berdasarkan keuntungan yang diperoleh musuh berdasarkan kebijakan serangan optimal.

Dalam skenario serangan kotak putih, kebijakan serangan dapat dirumuskan sebagai beberapa langkah mirip EM yang berupaya memperkirakan fungsi biaya musuh (jarang maupun padat) sebagai serangan kotak hitam dan kerugian pembelajar berfungsi sebagai serangan kotak putih. Memperkenalkan model pembelajaran metrik yang mengukur representasi gambar dan perbedaan informasi antara data sah dan tidak sah akan memungkinkan kita mengeksplorasi serangan substitusi kotak putih pada perilaku fungsi kerugian yang diharapkan dalam jaringan klasifikasi multi-label.

Perbandingan Antara Single Play Attack dan Multiple Play Attack pada Fungsi Custom Loss

Teknologi serangan permusuhan ada dalam visi komputer, pemrosesan bahasa alami, dan keamanan dunia maya pada data multidimensi, tekstual, dan gambar, data urutan, dan data spasial. Permasalahan tersebut mempelajari manipulasi fitur, biaya kesalahan klasifikasi, dan ketahanan distribusi dalam analisis malware, penyimpangan konsep, deteksi objek, deteksi kebaruan, deteksi outlier, deteksi peristiwa, klasifikasi tidak seimbang, pergeseran distribusi, penambangan pola langka, deteksi contoh di luar distribusi. , prediksi terstruktur, penambangan motif, kesalahan spesifikasi model, dan pembelajaran fitur non-stasioner.

Di sini, penelitian kami terhadap teori permainan mampu menghasilkan kriteria konvergensi kebijakan pencarian stokastik dan optima teoritis permainan dalam desain algoritma optimasi kuat berskala besar yang diperlukan dalam kecerdasan komputasi sistem cyber-fisik. Fungsi kerugian permusuhan dan prosedur pelatihan yang dihasilkan dalam penelitian kami dapat diterapkan untuk mempelajari kelayakan penerapan dan evaluasi pembelajaran mendalam.

Beruang dkk. membahas peran fungsi kerugian dalam menghargai akurasi dan menghukum ketidakakuratan. Fungsi kerugian cembung ternyata lebih menyukai model sederhana yang memiliki lebih banyak bias dan lebih sedikit varians. Sebaliknya, fungsi kerugian cekung lebih menyukai model kompleks yang memiliki bias lebih sedikit dan varian lebih banyak. Pertukaran optimalitas antara bias dan varians dalam memprediksi label kelas target dan kelas konsep terkait membentuk lanskap fungsi tujuan dari fungsi kerugian adversarial yang menginformasikan inferensi statistik dalam pembelajaran mendalam adversarial teoretis permainan. Menghasilkan dan menjelaskan manipulasi adversarial mengharuskan kita mempelajari efek bias algoritmik dalam masalah pembelajaran mendalam dan optimasi kuat selanjutnya dalam masalah pembelajaran adversarial.

Dekomposisi bias-varians dalam fungsi imbalan adversarial dapat dianalisis untuk memperoleh batas utilitas pembelajaran mendalam dalam kerangka batas kesalahan untuk aplikasi keamanan siber dari pembelajaran adversarial. Kami kemudian dapat mengungkapkan trade-off bias-varians dalam mempelajari ketahanan, keadilan, dan transparansi dalam kerangka pembelajaran mendalam untuk kecerdasan buatan yang dapat dijelaskan. Di sini, kita dapat mengeksplorasi pemfilteran, deteksi, dan estimasi sinyal dalam tensor dengan mekanisme divergensi informasi yang dibangun berdasarkan manipulasi persaingan teoritis permainan. Domingos menyajikan dekomposisi bias-varians terpadu yang berlaku untuk kerugian kuadrat, kerugian nol-satu, biaya kesalahan klasifikasi variabel, dan fungsi kerugian permusuhan. Ini digunakan dalam desain pembelajaran pohon keputusan, pembelajaran berbasis contoh, dan peningkatan.

Fungsi kerugian didefinisikan sebagai fungsi matematika untuk mengukur biaya model dalam memprediksi label klasifikasi atau nilai regresi. Tujuan mesin dinyatakan sebagai menghasilkan model dengan kerugian sekecil mungkin. Model optimal diperoleh dengan meminimalkan kerugian yang diharapkan pada semua contoh pelatihan, contoh validasi, dan contoh permusuhan. Dalam kasus kerugian nol-satu, model optimalnya adalah pengklasifikasi Bayes dengan fungsi kerugian yang disebut laju Bayes. Karena kerugian adalah fungsi dari kumpulan data pelatihan, algoritme pembelajaran adversarial yang sama menghasilkan model pembelajaran mesin yang berbeda untuk kumpulan data pelatihan yang berbeda. Ketergantungan ini dikurangi dengan merata-ratakan kerugian yang diperkirakan pada beberapa set data pelatihan yang mencakup set data pembelajaran adversarial.

Di sini, dekomposisi bias-varians menguraikan kerugian yang diharapkan menjadi bias, varians, istilah noise yang dihitung dengan algoritma komputasi. Distribusi data yang berlawanan diperhitungkan dalam istilah kebisingan. Istilah bias tidak bergantung pada rangkaian pelatihan dan bernilai nol bagi pelajar yang selalu membuat prediksi optimal. Istilah varians tidak bergantung pada nilai sebenarnya dari variabel yang diprediksi. Tidak ada gunanya bagi pembelajar yang selalu membuat prediksi yang sama apa pun rangkaian pelatihannya. Distribusi margin untuk mengklasifikasikan prediksi dengan benar dengan keyakinan tinggi kemudian dapat digunakan untuk mendapatkan batasan kesalahan generalisasi fungsi kerugian adversarial yang diusulkan dalam pembelajaran mendalam adversarial teoretis permainan.

Semakin kecil kemungkinan margin yang lebih rendah, semakin rendah batasan kesalahan generalisasi pada contoh pelatihan yang ditambah dengan contoh permusuhan. Memaksimalkan margin klasifikasi dan meminimalkan kesalahan klasifikasi merupakan kombinasi dari pengurangan jumlah contoh yang bias, penurunan varian model pada contoh yang tidak bias, dan peningkatan varians model pada contoh yang bias. Pekerjaan terkait adalah teori pembelajaran mendalam permusuhan dalam pola penambangan data dan teori pembelajaran mesin dalam algoritma pembelajaran komputasi. Ini memiliki aplikasi dalam analisis malware, penambangan agen, kontrol cerdas, dan analisis risiko dunia maya dalam pemodelan kepercayaan atas keamanan dan privasi pembelajaran mesin.

Belkin dkk. mempelajari kapasitas model interpolasi jaringan saraf dengan rezim pelatihan kurva kinerja penurunan ganda yang menggabungkan praktik konvensional kurva risiko bias-varians berbentuk U untuk menyeimbangkan underfitting dan overfitting sesuai dengan minimalisasi risiko empiris. Kesederhanaan prediktor jaringan saraf didefinisikan pada kelas fungsi yang berisi fungsi interpolasi dengan keteraturan atau kelancaran karena bias induktif yang lebih sedikit sebagaimana diukur dengan norma ruang fungsi. Fungsi interpolasi dengan norma yang lebih kecil dianggap lebih sederhana. Fungsi biaya adversarial pada prediktor jaringan saraf dalam pengaturan pembelajaran mendalam adversarial bertindak sebagai ukuran regularisasi pada bias induktif dalam pembelajaran mendalam adversarial teoretis permainan. Di sini, teori margin adalah upaya terkait untuk menemukan kelas fungsi dalam pengklasifikasi adversarial.

Penelitian tentang optimalitas prediktor interpolasi diperlukan untuk memperluas kelas fungsi tersebut dalam regresi adversarial yang didekati dengan pengklasifikasi multi-label dengan keluaran bernilai vektor dan jumlah kerugian kuadrat pada setiap keluaran. Menggabungkan rezim interpolasi dan analisis data empirisnya dalam pembelajaran mendalam yang bermusuhan membuka jalur penelitian komputasi, statistik, dan matematis baru ke dalam sifat optimalitas dan batasan utilitas dari prediktor pembelajaran mendalam. Li dkk. menyelidiki dimensi ruang parameter untuk memecahkan masalah komputasi dengan jaringan saraf untuk pengawasan, penguatan, dan jenis pembelajaran lainnya. Hasil tersebut berguna untuk menemukan struktur lanskap objektif dalam pembelajaran mendalam permusuhan dengan representasi terkompresi dari jaringan saraf dalam dalam pengoptimalan kotak hitam.

Strumbelj dkk. menggunakan teori permainan koalisi untuk menjelaskan prediksi individu model klasifikasi. Metode penjelasan yang diusulkan dirancang untuk bekerja dengan semua jenis pengklasifikasi. Dalam pembelajaran mesin, hal ini dapat dibandingkan dengan metode penjelasan khusus model seperti aturan keputusan dan jaringan Bayesian serta metode yang memberikan penjelasan dalam bentuk kontribusi fitur dalam ansambel pengklasifikasi seperti hutan acak. Dalam pembelajaran mendalam, hal ini dapat dibandingkan dengan metode ekstraksi aturan yang diterapkan pada jaringan saraf untuk mengurangi ketergantungan antara kebutuhan pengguna akhir (yang diperoleh dari pemasaran, kedokteran, dll.) dan metode pembelajaran mesin yang mendasarinya.

Gagasan tentang perbedaan prediksi diajukan antara prediksi saat ini dan prediksi yang diharapkan sehubungan dengan kontribusi nilai fitur saat ini terhadap prediksi tersebut. Tidak ada asumsi yang dibuat mengenai relevansi sebelumnya dari nilai fitur individual. Perubahan dalam prediksi pengklasifikasi didekomposisi menjadi kontribusi fitur individu menggunakan konsep dalam teori permainan koalisi. Narayanam dkk. mengusulkan untuk menemukan node berpengaruh yang bertindak sebagai fitur yang dipelajari dalam jaringan sosial dengan nilai-nilai Shapley dan teori permainan kooperatif. Nilai-nilai Shapley adalah konsep solusi yang memberikan alokasi hasil yang diharapkan dalam desain permainan koalisi.

Masalah penyebaran informasi di jejaring sosial ditujukan untuk aplikasi seperti pemasaran viral, promosi penjualan, dan tren penelitian di jaringan penulisan bersama untuk ide-ide abstrak dan informasi teknis dengan algoritma komputasi yang efisien. Masalah pemilihan kumpulan node target untuk menemukan node yang berpengaruh dirumuskan sebagai masalah penemuan pola cakupan dalam penambangan data yang memodelkan keputusan individu yang dipengaruhi oleh perilaku tetangga terdekat di jaringan sosial. Konsep solusi nilai Shapley memenuhi sifat matematika yang disebut linearitas, simetri, dan sifat pembawa untuk menemukan cara yang adil dalam mendistribusikan keuntungan kerja sama di antara para pemain dalam permainan koalisi.

Nilai-nilai Shapley memperhitungkan semua kemungkinan dinamika koalisi dan skenario negosiasi di antara para pemain. Node dalam jaringan sosial dapat dianggap sebagai entitas individu yang berperilaku strategis dan berkepentingan dalam suatu organisasi yang berfungsi sesuai dengan desain mekanisme dalam teori permainan. Kemungkinan suatu node dipengaruhi oleh tetangganya tidak hanya bergantung pada struktur komunitas jaringan sosial tetapi juga informasi pribadi yang dimiliki node tersebut tentang tetangganya. Masalah kumpulan node target dapat digunakan untuk aplikasi pembelajaran mesin di bidang pemasaran, politik, ekonomi, epidemiologi, sosiologi, jaringan komputer, dan database.

Dalam penelitian selanjutnya, kami akan mengeksplorasi ketergantungan antara pengacakan dalam manipulasi permusuhan dan optimalisasi dalam formulasi permainan kami. Saat ini, permainan teoritis stokastik optima (pemecahan data adversarial) ditentukan oleh konvergensi fungsi biaya adversarial daripada fungsi biaya klasifikasi. Namun, fungsi biaya klasifikasi multi-label dalam ruang strategi multipemain yang terdiri dari strategi murni dan juga strategi campuran dapat menjadi jalur penelitian yang bermanfaat. Dengan asumsi skenario serangan kotak putih pada pengklasifikasi CNN berarti kita dapat memandu pengaturan parameter dalam algoritme genetika dan algoritme SA ke dalam distribusi data permusuhan yang bergantung pada aplikasi.

Skenario serangan dengan interaksi antara banyak musuh yang bekerja sama memunculkan permainan koalisi dalam gaya satu pemimpin-banyak pengikut dengan tujuan permusuhan tunggal atau ganda. Kami juga tertarik dengan strategi pengacakan untuk pengoptimalan yang kuat dalam game multipemain yang dapat diuraikan menjadi game prediksi atau game Stackelberg. Di sini, beberapa rumusan teori permainan yang relevan dapat ditemukan dalam literatur tentang permainan evolusi, permainan matriks, permainan kuat, permainan fuzzy, permainan Markov, dan permainan Bayesian. Dimensi data multi-label dalam jenis permainan ini dapat diatasi dengan melatih algoritma adversarial untuk aliran data dan paralelisasi aliran kontrol dengan beberapa unit pemrosesan. Operator pencarian terpandu dalam pembelajaran evolusi mungkin mengarah pada skenario serangan dengan paralelisasi untuk optimasi stokastik dalam estimasi besaran langkah dan arah permusuhan.

Selanjutnya, kita juga dapat bereksperimen dengan berbagai fungsi klasifikasi dengan mengubah arsitektur pembelajaran mendalam atau bereksperimen dengan metode optimasi multi-objektif dengan mengubah operator evolusi dalam evaluasi fungsi kebugaran optimasi stokastik dan permainan yang didorong oleh kendala. Ada kemungkinan bahwa musuh

variasional kami dengan model campuran Gaussian dapat ditingkatkan dengan model probabilistik yang disesuaikan, seperti model campuran multinomial dan jaringan kepadatan campuran untuk representasi data gambar. Kriteria validasi pengguna atas data permusuhan yang dihasilkan dapat diwakili oleh jaringan dalam seperti jaringan terjemahan spasial.

Kami juga berencana untuk menyelidiki skenario permainan multipemain yang lebih menantang di mana musuh menyerang beberapa label secara bersamaan. Dalam masalah pembelajaran permusuhan multipemain ini, kami ingin mensimulasikan manipulasi yang mengubah label positif yang ditargetkan menjadi salah satu dari banyak label negatif. Skenario serangan yang berhasil dalam ruang strategi jenis ini kemudian akan menginformasikan permainan multipemain dengan strategi campuran yang memiliki dua atau lebih label sebagai target manipulasi yang diberikan kepada satu pembelajar. Dalam skenario ini, pengacakan strategi dan hasil dalam formulasi permainan kemungkinan besar akan memengaruhi regularisasi bobot dan batasan keputusan pelajar.

Kita juga dapat menerapkan teori permainan dan teori kontrol untuk mengoptimalkan pemodelan numerik dalam kerangka pembelajaran mendalam permusuhan dari skenario serangan berulang dan optimalisasi pertahanan untuk deteksi dinamika, karakterisasi, dan prediksi dalam pemilihan fitur. Label kelas dan representasi pengetahuan harus dihasilkan untuk objek yang tidak diketahui yang diklasifikasikan sebagai prediksi multimodal, multi-tampilan, dan multitugas. Mereka akan mengarahkan kita untuk menggabungkan inferensi statistik dalam prosedur pelatihan jaringan saraf dalam dengan desain pengklasifikasi yang hemat biaya dan pengoptimalan yang kuat dalam deteksi dinamika kompleks. Fungsi kerugian khusus dalam pembelajaran mendalam dapat dirancang untuk memeriksa ukuran ketahanan distribusi dalam pemilihan fitur permusuhan.

Untuk menerapkan inferensi probabilistik dalam keamanan siber, kami akan merancang fungsi kerugian khusus untuk klasifikasi asosiatif dengan estimasi kemungkinan maksimum (MLE). Metode klasifikasi asosiatif digunakan untuk penemuan aturan, pemeringkatan aturan, pemangkasan aturan, prediksi aturan, dan evaluasi aturan dalam penambangan pola frekuensi. Dalam metode ini, pola frekuensi yang signifikan secara statistik adalah pola yang dianggap informatif dan tidak berlebihan menurut ukuran ketertarikan. Ukuran ketertarikan yang populer berusaha mengendalikan laju penemuan pola frekuensi yang salah dengan menghitung frekuensinya.

Definisi frekuensi umum mencakup dukungan, kepercayaan diri, peningkatan, pengaruh, keyakinan, dan peningkatan. Pada penelitian sebelumnya, kami mengamati bahwa risiko penemuan palsu dapat dinilai dengan menerapkan model pembelajaran diskriminatif untuk menemukan pola yang sering terjadi. Meminimalkan kesalahan diskriminasi memungkinkan kita memperoleh pola frekuensi yang signifikan secara statistik untuk menilai fungsi kerugian teoretis informasi yang ditentukan dalam catatan data pelatihan. Kami juga telah menggeneralisasi ukuran ketertarikan dari pola yang sering ditentukan pada sampel data masukan ke ekspektasi kondisional yang diperkirakan terhadap populasi data dalam database yang mendasarinya.

Ekspektasi bersyarat menentukan masalah optimasi cembung dalam hal matriks kejadian fungsi indikator yang menghubungkan catatan data dengan pola yang sering terjadi dalam pemodelan MLE. Kami kemudian mengusulkan aturan pembaruan dan aturan keputusan untuk memecahkan masalah optimasi cembung dengan algoritma penskalaan berulang (IS). Kemudian kompleksitas komputasi algoritma IS ditentukan oleh kepadatan matriks kejadian. Dalam konteks ini, kita akan mengintegrasikan penambangan pola frekuensi dengan pembelajaran mendalam adversarial untuk menemukan representasi ringkas dari pola frekuensi yang paling informatif menurut teori informasi. Kami juga akan memeriksa inferensi bersyarat dan pemodelan variasional dari pola yang sering terjadi dimana fungsi indikator dibatasi oleh fitur fuzzy.

Kami akan mensurvei alat komputasi dalam statistik multivariat yang mempelajari sistem distribusi probabilitas. Kemudian kita dapat merancang fungsi kerugian khusus dalam pembelajaran mendalam dengan tujuan teoretis permainan. Sebelumnya, kami telah mengerjakan pembelajaran fitur kausal Granger dalam prediksi deret waktu multivariat yang dihasilkan oleh jaringan regresi mendalam untuk penambangan data. Mereka mempelajari risiko empiris dari sistem yang kompleks dengan fungsi kerugian khusus. Fitur kausal Granger dalam proposal kami meningkatkan kesalahan regresi multivariat dengan pembelajaran mendalam. Pada saat yang sama, pemahaman yang dapat dijelaskan tentang distribusi data masukan ditingkatkan dengan hubungan sebab-akibat deskriptif yang lebih informatif daripada koefisien korelasi dan bobot jaringan saraf untuk menjelaskan hasil regresi.

Di sini, fungsi pembayaran teoritis game mengukur optimalisasi yang didorong oleh pemain yang meningkatkan pelatihan dan inferensi dalam pembelajaran mesin dan lingkungan yang tidak pasti. Mereka juga menjelaskan dampak lingkungan yang tidak pasti dengan mengacu pada distribusi hasil, dan, dalam arti rasionalitas teori keputusan, fungsi pembayaran memaksimalkan utilitas yang diharapkan untuk setiap pemain dalam permainan.

Mesin Paralel dalam Game Tereduksi

Cai dkk. membahas keseimbangan teoritis permainan dalam permainan dua pemain bukan jumlah nol dengan generalisasi multipemain. Teorema min-maks terbukti untuk permainan zero-sum polimatrix multipemain. Ekuilibrium Nash ditemukan dengan program linier. Permainan polimatriks ditentukan oleh graf yang simpulnya adalah pemain dengan strategi terkait dan sisinya adalah permainan dua pemain. Imbalan seorang pemain adalah jumlah seluruh imbalan dalam permainan yang berdekatan dengannya. Permainan polimatriks jumlah nol mewakili sistem pembayaran tertutup.

Strategi keseimbangan untuk permainan ini adalah strategi maksimal yang mewakili permainan tanpa penyesalan untuk semua pemain. Oliehoek dkk. menyajikan permainan asimetris untuk pencarian dalam algoritma ko-evolusi yang tidak memerlukan spesifikasi fungsi evaluasi. Dalam permainan asimetris, strategi pemain saat ini ditentukan oleh tindakan yang diambil oleh pemain sebelumnya. Algoritme ko-evolusi seperti ini berguna dalam permasalahan algoritmik seperti pembelajaran mesin teori permainan, pembelajaran konsep, jaringan pengurutan, klasifikasi kepadatan menggunakan automata seluler, serta perkiraan dan klasifikasi fungsi. Kasus evaluasi yang kompleks dapat dibangun dengan proses pencarian.

Strategi berkualitas tinggi dikembangkan selama pencarian. Konsep solusi dalam proses pencarian menentukan kandidat solusi mana yang memenuhi syarat sebagai solusi optimal dan mana yang tidak. Algoritme pembelajaran teoretis permainan merancang kriteria konvergensi pada utilitas yang diharapkan yang ditentukan oleh konsep solusi yang dioptimalkan terhadap musuh yang cerdas. Ekuilibrium Nash kemudian menetapkan strategi campuran (acak) untuk setiap pemain yang tidak memiliki insentif untuk menyimpang berdasarkan strategi pemain lain. Jadi solusi teoritis permainan untuk masalah pembelajaran mesin adalah rekomendasi permainan yang optimal untuk semua pemain. Ini mengarah pada sistem multi-agen dengan keseimbangan Nash sebagai konsep solusi.

Mereka juga dapat memasukkan ko-evolusi Pareto untuk mengakomodasi banyak musuh dengan tujuan berbeda. Strategi respons terbaik diselesaikan dengan proses pengambilan keputusan Markov yang dapat diamati sebagian, sesuai dengan permainan asimetris bentuk ekstensif terbatas yang disebut memori Nash paralel. Semua kemungkinan keyakinan pemain dan transisi di antara mereka dapat dihasilkan dalam proses pengambilan keputusan Markov. Metode maksimalisasi bergantian atau optimasi pendakian koordinat memecahkan strategi respons terbaik.

Bianchi dkk. mensurvei kerangka algoritmik yang disebut metaheuristik untuk memecahkan masalah optimasi kompleks dalam pembelajaran mendalam permusuhan teoritis permainan dengan formulasi matematika untuk informasi yang tidak pasti, stokastik, dan dinamis. Optimalisasi koloni semut, komputasi evolusioner, simulasi anil, dan pencarian tabu semuanya merupakan metaheuristik yang dapat diterapkan pada masalah optimasi kombinatorial stokastik (SCOPs) dalam menghasilkan manipulasi permusuhan di setiap iterasi permainan permusuhan. Dalam pemecahan masalah dengan SCOPs, informasi tentang data masalah sebagian tidak diketahui seperti informasi yang tersedia bagi pelajar tentang strategi musuh tertentu.

Selain itu, SCOP mengasumsikan distribusi probabilitas tentang pengetahuan tentang data masalah seperti karakterisasi tipe musuh. SCOP diselesaikan dengan pemrograman dinamis. Metaheuristik tersebut menggabungkan beberapa heuristik yang merupakan algoritma pencarian lokal yang dimulai dari solusi/pergerakan yang sudah ada sebelumnya untuk fitur yang dipelajari atau algoritma konstruktif yang melakukan konstruksi fitur/komponen dari solusi untuk menemukan manipulasi permusuhan dalam pembelajaran mendalam permusuhan teoritis permainan. Dibandingkan dengan heuristik, metaheuristik mencapai keseimbangan stokastik dinamis antara mengeksploitasi ruang pencarian yang mewakili akumulasi pengalaman secara efektif dan secara efisien menjelajahi wilayah baru dari ruang pencarian dengan solusi berkualitas tinggi.

Bukti konvergensi untuk metaheuristik berguna untuk memperoleh wawasan analitik tentang prinsip kerja algoritma komputasi. Namun, mereka mengasumsikan waktu komputasi, ruang memori, dan ukuran sampel yang tak terbatas berguna dalam penerapan metaheuristik yang efisien dalam praktiknya. Pemrograman bilangan bulat stokastik, pemrograman dinamis stokastik, optimasi simulasi, metode partisi stokastik, lindung nilai progresif, dan pencarian lingkungan variabel adalah area penelitian di SCOPs. Mereka dapat

digunakan untuk memperluas penelitian kami dalam pembelajaran mendalam permainan teoritis permusuhan ke simulasi dunia nyata yang kompleks yang menjembatani kesenjangan antara teori dan praktik.

Martin memperkenalkan metaheuristik untuk menggabungkan metode anil simulasi stokastik dengan metode pencarian lokal deterministik untuk menghasilkan rantai Markov baru untuk optimasi global. Metaheuristik baru diuji pada masalah optimasi kombinatorial (CO) seperti masalah travelling salesman dan masalah partisi grafik yang harus berurusan dengan ukuran data yang besar. Iterasi rantai Markov disebut optimasi lokal berantai yang bertindak sebagai generalisasi aturan pembaruan dalam simulasi anil. Ini dijalankan secara paralel pada jaringan lokal workstation. Arsitektur memori terdistribusi dan sistem penyampaian pesan menjalankan rantai Markov secara simultan untuk menghasilkan populasi kandidat solusi dengan melakukan pencarian independen untuk pergerakan yang menguntungkan.

Percabangan dan pemangkasan rantai Markov dibagi di antara populasi yang dihasilkan untuk menduplikasi kandidat terbaik dengan mengorbankan kandidat terburuk. Pencarian paralel secara dinamis beradaptasi untuk memberikan pencarian terbanyak pada prosesor tercepat. Suman dkk. meninjau algoritma simulasi anil untuk masalah optimasi tujuan tunggal dan multi-tujuan untuk mendapatkan solusi optimal dan himpunan solusi optimal Pareto. Anil simulasi Pareto dirangkum. Ini menciptakan populasi sampel dari solusi yang berinteraksi untuk menghasilkan perkiraan yang baik terhadap solusi efisien dengan konsep lingkungan yang secara probabilistik menerima solusi baru.

Anil simulasi Pareto dapat digunakan untuk menciptakan solusi efisien terhadap manipulasi permusuhan dalam pembelajaran mendalam permusuhan teoritis permainan. Anil simulasi dapat digunakan dalam optimalisasi tujuan desain yang berganda dan saling bertentangan dalam permainan permusuhan multipemain dengan fungsi biaya multimodal dan tidak lancar dalam pemrosesan sinyal permusuhan. Algoritme anil simulasi multi-tujuan dapat mengakomodasi penanganan kendala dalam pemecahan masalah praktis. Selain itu, penerapan simulasi anil pada pengenalan pola hibrid dan klasifikasi objek juga dimungkinkan. Rajasekaran dkk. membahas konsep simulasi anil sebagai keluarga algoritma acak.

Bukti konvergensi disediakan untuk menerapkan simulasi anil dalam masalah optimasi dengan sifat teoritis grafik khusus pada fungsi biaya. Algoritma nested annealing dikembangkan untuk menentukan grafik pencarian yang sesuai dengan masalah optimasi yang diberikan. Kriteria “keterpisahan” yang bergantung pada aplikasi dari grafik pencarian ke dalam subgraf dianalisis untuk memberikan batasan ketat pada perilaku yang diharapkan dari anil bersarang. Solusi konvergen mewakili vektor keadaan probabilitas dari rantai Markov yang mewakili algoritma anil bersarang.

Fogel dkk. membahas simulasi anil untuk optimasi stokastik dalam metode simulasi evolusi. Evolusi yang disimulasikan selanjutnya dikategorikan ke dalam algoritma genetika, strategi evolusi, dan pemrograman evolusioner dan ditinjau secara komprehensif. Implementasinya pada mesin paralel dan arsitektur pemrosesan terdistribusi juga dibahas. Pilihan antara fungsi tujuan, fungsi probabilitas pembangkitan, fungsi probabilitas

penerimaan, jadwal pendinginan, dan lingkungan pencarian diberikan sehubungan dengan masalah optimasi di beberapa domain aplikasi.

Henderson dkk. membahas teori optimasi cybernetic dan praktik gerakan mendaki bukit dalam algoritma simulasi anil. Simulasi anil dibandingkan dan dikontraskan dengan pencarian tabu dan algoritma genetika. Ram dkk. mengembangkan algoritma anil simulasi paralel untuk masalah optimasi non-linier yang kompleks. Algoritma anil simulasi paralel dikategorikan menjadi paralelisme percobaan tunggal dan paralelisme percobaan ganda. Mereka juga dikategorikan ke dalam algoritma mirip serial, algoritma yang dihasilkan diubah, dan algoritma asynchronous berdasarkan trade-off antara akurasi fungsi biaya, pembangkitan keadaan, paralelisme, dan overhead komunikasi. Arsitektur komputer tujuan khusus dapat dirancang untuk mengimbangi pengorbanan yang terlibat secara komputasi dalam algoritma anil.

Untuk pemfilteran dan estimasi, pengelompokan, dan klasifikasi objek prediksi pengenalan pola dalam proses stokastik spatio-temporal yang menghasilkan fitur adversarial, kita harus merumuskan pembelajaran representasi mendalam untuk pembelajaran adversarial teoretis permainan. Dalam konteks ini, kita dapat mengeksplorasi teori pembelajaran adversarial dengan permainan tereduksi, fungsi submodular, analisis wavelet, dan aplikasi pelatihan adversarial dengan deteksi objek yang kuat, mesin faktorisasi, pembelajaran kamus, dan komputasi granular. Di sini, pembelajaran kamus menghasilkan kamus elemen filter untuk merekonstruksi representasi data pelatihan yang sangat berlebihan dengan model pengkodean yang jarang dalam masalah pengoptimalan dan inferensi berbasis data untuk pembelajaran mendalam yang bermusuhan.

Di sini, mesin faktorisasi adalah perkiraan tingkat rendah dari rekayasa fitur tensor data renggang ketika sebagian besar elemen prediksinya tidak diketahui. Di sini, komputasi granular berguna untuk membuat aturan fusi data pada representasi fitur data pelatihan. Hal ini dapat menyebabkan sistem neuro-fuzzy dan sistem multi-agen dalam penambangan data. Kita dapat menyelidiki lebih lanjut transfer aturan fusi data yang signifikan secara statistik antara representasi data prediktif pada resolusi spasial dan distribusi data resolusi spektral dari manifold pelatihan. Di sini, pembelajaran kamus adalah paradigma pembelajaran mesin komputasi yang dapat menganalisis pembuatan fitur multimodal dan masalah pengoptimalan multivariat dalam tugas prediksi.

Pengacakan, konvergensi, dan paralelisasi yang lebih baik pada besaran langkah algoritma optimasi teori permainan akan menghasilkan kebijakan stokastik yang lebih baik dalam keseimbangan teori permainan. Di sini, permainan Stackelberg dengan fungsi hasil permusuhan optimal akan menghasilkan keseimbangan Nash dan Stackelberg serta Pareto optima dalam ruang strategi acak permainan. Minat penelitian kami dalam penambangan data adalah pada pemodelan prediktif dengan permainan stokastik dan pembelajaran mendalam permusuhan. Metode penambangan data kami kemudian mampu mengakomodasi tantangan baru pada metodologi analitik yang diterapkan oleh model data besar, di mana biasanya tidak mungkin untuk menyimpan seluruh aliran data berdimensi

tinggi atau memindainya berkali-kali karena volumenya yang sangat besar dan dinamika yang berubah-ubah.

Distribusi data yang mendasarinya dari waktu ke waktu. Di sini, algoritma paralel dan komputasi terdistribusi juga akan menghasilkan pengurangan yang signifikan dalam biaya komputasi diamortisasi dari penambangan pola permusuhan pada data berdimensi tinggi dan multidimensi. Pilihan model pemrograman pada mesin paralel adalah antara paralelisme data, paralelisme tugas, dan paralelisme grafik. Detail penerapan analisis data terdistribusi dalam pembelajaran mendalam adversarial harus mempertimbangkan kerangka pengembangan penambangan pola untuk analisis data serial pada data besar. Ini mencakup asumsi pembelajaran mesin pada model data, model memori, model pemrograman, model komunikasi, model eksekusi, dan model komputasi untuk mengubah algoritma serial menjadi algoritma paralel dalam pembelajaran mendalam permusuhan.

Arsitektur jaringan dalam hibrid dapat diusulkan untuk komposisi semantik pada urutan peristiwa masukan dalam semiotika pembelajaran mendalam. Di sini, kita perlu mempelajari representasi terdistribusi dari data multi-relasional yang diambil dari basis pengetahuan. Kami berencana merancang model pembelajaran tanpa pengawasan untuk penambangan motif dengan pengelompokan biclustering dan evolusioner, pengelompokan bertingkat, dan pengelompokan berbasis model. Untuk menciptakan pembelajaran terawasi dengan motif seperti itu, kami akan fokus pada metode kompresi dan metode optimasi dalam pembelajaran kernel dan pembelajaran mendalam.

Teori yang relevan dalam penambangan data adalah pengelompokan bertingkat, partisi grafik bertingkat, deteksi kuasi-klik, dan penemuan subgraf padat. Kami dapat menskalakan pembelajaran mendalam permusuhan teoretis permainan tersebut ke pengaturan data besar dengan metode pengambilan sampel data yang dapat mengatasi dimensi data dan granularitas data untuk multiprosesor dan pemrosesan batch paralel yang memalukan melalui tensor dan grafik. Pekerjaan penelitian terkait adalah studi tentang metode pengambilan sampel seperti undersampling, oversampling, ketidakpastian sampling, reservoir sampling, struktural sampling, dll. Solusi data besar akan melibatkan operasi rekayasa data untuk caching, pengurutan, pengindeksan, hashing, pengkodean, pencarian, partisi, pengambilan sampel, dan pengambilan dalam model inkremental, model urutan, dan model ansambel untuk pembelajaran permusuhan yang sensitif terhadap biaya dengan model probabilistik.

4.4 PERMAINAN STOKASTIK DALAM PEMODELAN PREDIKTIF

Interaksi antara musuh dan pengklasifikasi telah dimodelkan sebagai permainan Stackelberg. Di sini, peran musuh bukanlah sebagai penghasil data statis, melainkan sebagai agen cerdas yang sengaja melakukan manipulasi data untuk menghindari pengklasifikasi. Kegagalan dalam mempertimbangkan penghindaran permusuhan dalam desain pengklasifikasi menimbulkan masalah keamanan dalam deteksi penipuan, deteksi intrusi komputer, pencarian web, deteksi spam, dan aplikasi deteksi phishing. Mempelajari kembali bobot pengklasifikasi adalah solusi yang lemah untuk klasifikasi yang kuat karena serangan

penghindaran dihasilkan dengan kecepatan yang lebih murah dan lebih cepat daripada pembelajaran ulang.

Li dkk. mengusulkan serangan substitusi silang fitur untuk menunjukkan musuh yang didorong oleh tujuan mengeksploitasi keterbatasan pengurangan fitur dalam pengaturan permusuhan. Musuh dapat menanyakan pengklasifikasi berdasarkan anggaran kueri tetap dan anggaran biaya tetap. Model penghindaran permusuhan dengan pengatur renggang kemudian disajikan. Membangun pengklasifikasi pada kelas kesetaraan fitur daripada ruang fitur diusulkan sebagai solusi untuk meningkatkan ketahanan pengklasifikasi. Solusi lain mengusulkan permainan interaksi bilevel Stackelberg antara pengklasifikasi dan kumpulan musuh. Permainan Stackelberg diselesaikan dengan pemrograman linier bilangan bulat campuran dengan pembangkitan kendala.

Bianchi dkk. menyajikan permainan berulang untuk masalah prediksi acak. Masalah prediksi sekuensial dimodelkan dalam kerangka keseimbangan Nash yang ditemukan dalam permainan bentuk normal. Teorema min-maks tertentu dibahas untuk menganalisis permainan zero-sum dua pemain. Kerangka batas kesalahan disediakan untuk menganalisis algoritma pembelajaran teoritis permainan. Bruckner mengusulkan permainan prediksi untuk memodelkan interaksi antara pelajar yang membangun model prediktif dan musuh yang mengontrol proses pembuatan data.

Kerangka permainan prediksi memungkinkan model eksplisit tentang minat, tindakan, pengetahuan, dan keputusan pemain teoretis permainan. Kemudian kesalahan generalisasi model prediktif dianalisis dalam istilah ekuilibrium Nash dengan menyelesaikan permainan Stackelberg statis dan dinamis dua pemain non-kooperatif. Permainan zero-sum stokastik dua pemain yang menggabungkan banyak musuh dianalisis oleh Ummels. Cai dkk. menganalisis game polimatriks yang memberikan generalisasi multipemain ke game zero-sum dua pemain. Strategi ekuilibrium Nash yang sesuai mungkin menyimpang dari optimasi max-min dan diperoleh dari solusi program linier.

Zhou dkk. mensurvei permainan Stackelberg dua pemain dan banyak pemain dalam algoritma pembelajaran permusuhan dan aplikasi keamanan siber. Interaksi antara musuh dan pengklasifikasi dimodelkan sebagai satu atau lebih permainan simultan dan permainan berurutan dimana musuh dapat menjadi pemimpin atau pengikut dalam permainan. Alpcan dkk. menyajikan permainan strategis berskala besar dan pengurangan konsumsi komputasi permainan dan batasan informasi pada solusi keseimbangan Nash.

Oliehoek dkk. mengusulkan musuh generatif yang mendalam dengan respons terbaik yang dibatasi sumber daya dan keseimbangan Nash pada data sintetis. Musuh generatif memiliki jaringan generator dan jaringan diskriminator dalam masalah pembelajaran yang diawasi dan dioperasikan pada data diskrit. Secara khusus, fungsi kerugian jaringan generator hanya bergantung pada data “palsu”, sedangkan fungsi kerugian jaringan diskriminator bergantung pada data “asli” dan data “palsu”. Baik generator maupun diskriminator berpartisipasi dalam permainan bentuk strategis zero-sum di mana fungsi pembayaran masing-masing pemain ditentukan pada ruang strategi campuran.

Papernot dkk. memberikan model ancaman yang merangkum berbagai skenario serangan dalam algoritma pembelajaran permusuhan. Ancaman terhadap model pembelajaran mesin adalah manipulasi permusuhan, yang dihasilkan selama proses pelatihan dan proses inferensi. Dalam tahap pelatihan, musuh dapat memanipulasi proses pengumpulan data dengan memasukkan contoh permusuhan ke dalam data pelatihan dengan tujuan untuk mengubah batasan keputusan model pembelajaran. Pada tahap inferensi, musuh dapat merencanakan serangan black-box atau white-box pada parameter model pembelajaran untuk menyebabkan penyimpangan distribusi antara distribusi data pelatihan dan distribusi data runtime. Papernot dkk. juga mengusulkan teorema tidak ada makan siang gratis dan mungkin model yang kira-kira tepat untuk pembelajaran permusuhan. Yang dkk. menganalisis keseimbangan Nash dari sistem dinamik diferensial yang memodelkan skenario serangan siber ancaman persisten tingkat lanjut (APT).

Bowling dkk. membahas pembelajaran kebijakan di lingkungan multi-agen menggunakan permainan stokastik. Rasionalitas dan konvergensi agen pembelajaran dirumuskan untuk merancang algoritma kebijakan pendakian bukit. Masalah pembelajaran multi-agen merancang algoritma pembelajaran untuk agen yang diminati dengan adanya agen pembelajaran lain di luar kendalinya. Pembelajaran seperti ini dianggap memiliki target bergerak menuju kebijakan optimal yang harus mempertimbangkan perubahan adaptasi agen lain, selama agen pembelajaran tersebut melakukan pelatihan.

Di sini, permainan stokastik adalah perpanjangan multi-agen dari proses keputusan Markov (MDP) yang cocok untuk pembelajaran agen tunggal seperti Q-learning. Permainan stokastik seperti itu mempunyai kegunaan dalam pembelajaran penguatan permusuhan. Hal ini dapat ditingkatkan dengan ide-ide dari teori permainan evolusioner seperti dinamika replikator yang disesuaikan, dan algoritma pembobotan acak dapat diperkenalkan ke dalam pembelajaran mendalam adversarial teoretis permainan untuk mendistribusikan ulang bobot dengan permainan multipemain.

Bowling dkk. mengembangkan perilaku pembelajaran sistem multi-agen yang beradaptasi dengan lingkungan non-stasioner. Teknik yang relevan untuk menyelesaikan permainan stokastik untuk menemukan keseimbangan diberikan dari referensi dalam bidang penelitian teori permainan dan pembelajaran penguatan. Persamaan dan perbedaan antara algoritma komputasi pada area penelitian ini diidentifikasi. Asumsi tentang kontrol dan batasan keyakinan algoritma komputasi tersebut dapat diterapkan ke dalam pembelajaran mendalam adversarial teoretis permainan. Permainan stokastik yang dihasilkan merupakan perpanjangan dari permainan matriks jumlah umum yang lebih sederhana yang diselesaikan dengan pemrograman kuadrat dalam teori permainan.

Dalam teori permainan, mereka mewakili interaksi antara pelajar dan lingkungannya untuk menemukan nilai keseimbangan dari formulasi permainan tetapi bukan kebijakan keseimbangan di bawah batasan fisik atau rasional. Sebaliknya, pembelajaran penguatan tidak mengasumsikan pengetahuan tentang lingkungan belajar. Agen pembelajaran penguatan bertindak hanya berdasarkan pengamatan transisi antara keadaan pembelajar dan fungsi penghargaan pada tindakan. Tujuan agen pembelajaran penguatan adalah

menemukan kebijakan optimal pada keseimbangan permainan. Menemukan konsep solusi untuk permainan stokastik yang memiliki pemodelan yang bergantung pada lawan juga memerlukan penyelidikan teknik generalisasi dan perkiraan pada kumpulan data kompleks yang memiliki asumsi dan batasan implisit untuk pembelajaran mendalam yang bermusuhan.

Kami mengusulkan respons terbaik non-cembung baru dalam setiap permainan penyelesaian permainan prediksi untuk manipulasi permusuhan. Masalah optimasi stokastik bilevel kami dalam permainan prediksi dirumuskan sebagai permainan Stackelberg dua pemain dengan jumlah variabel berurutan yang berulang. Masalah optimasi diselesaikan dengan prosedur pencarian Alternating Least Squares (ALS) yang terus-menerus menyerang pengklasifikasi yang dilatih ulang dengan manipulasi adversarial yang dioptimalkan hingga keseimbangan Nash. Prosedur ALS mengevaluasi manipulasi kandidat permusuhan yang dihasilkan oleh prosedur simulasi anil (SA) untuk peningkatan fungsi hasil permusuhan pada label kelas yang ditargetkan. Oleh karena itu, data permusuhan yang dihasilkan dalam permainan Stackelberg mensimulasikan interaksi berkelanjutan, bukan interaksi satu kali dengan proses pembelajaran pengklasifikasi.

Kerangka Teori Pembelajaran Komputasi untuk Menganalisis Algoritma Pembelajaran Teori Game

Johnson dkk. melakukan analisis ukuran perilaku kepercayaan dan dapat dipercaya dengan “permainan kepercayaan” dan “permainan investasi” dalam organisasi, ekonomi, dan masyarakat. Meningkatkan kepercayaan meningkatkan efisiensi dengan menurunkan biaya. Kesediaan bersama karena perilaku yang dapat dipercaya menghasilkan hasil ekonomi per kapita yang lebih baik. Kepercayaan yang terkumpul menghasilkan sistem peradilan yang efisien, birokrasi pemerintah yang lebih berkualitas, korupsi yang lebih rendah, dan pembangunan keuangan yang lebih besar. Model perilaku lainnya memasukkan preferensi keadilan ke dalam utilitas pemain. Di dalamnya, teori kesetaraan mengarah pada prediksi perbedaan perilaku ketika kekayaan tidak setara.

Permainan kepercayaan harus menghasilkan tingkat pengembalian kepercayaan yang positif sebagai perolehan kesejahteraan yang memfasilitasi pertukaran lebih lanjut. Wali amanat menanggapi setiap kemungkinan perilaku dari pihak lawan untuk menentukan perilaku yang adil. Lawannya bisa menjadi lawan yang disimulasikan dalam lingkungan yang bermusuhan. Proses memikirkan implikasi perilaku dari setiap kemungkinan hasil mengubah persepsi pemain terhadap permainan dan mengarah pada pengambilan keputusan berdasarkan data. Ketika jaminan anonimitas di antara para pemain dihilangkan, perilaku dimotivasi oleh faktor-faktor seperti kepedulian individu terhadap reputasi mereka, tindakan baik di masa lalu, dan ketakutan akan pembalasan. Kemungkinan keuntungan dalam pertukaran di masa depan antar pemain memotivasi imbalan yang diharapkan dari perilaku pemain lawan selain kepercayaan pada lawannya.

Statistik deskriptif untuk kepercayaan dan kelayakan dihasilkan mengikuti spesifikasi regresi kuadrat terkecil dari hasil permainan kepercayaan. Distribusi outlier dan adversarial dalam data evaluasi tersebut secara signifikan membuat bias estimasi parameter.

Pemeriksaan manipulasi permusuhan harus memverifikasi contoh-contoh permusuhan untuk membedakan antara simulasi dan manusia untuk menentukan hasil permainan kepercayaan. Di sini, kekuatan permusuhan dan efek interaksi yang ditemukan dari solusi pembelajaran mendalam permusuhan teoritis permainan menambah strategi pengambilan keputusan heuristik dan aturan praktis yang terkandung dalam budaya sebagai strategi untuk percaya atau tidak percaya.

Balduzzi dkk. menyajikan bahasa umum untuk mendeskripsikan dan menganalisis algoritmik teoretis permainan dalam pembelajaran mendalam. Pelatihan propagasi mundur jaringan saraf dalam dinyatakan dalam bentuk optimasi terdistribusi dalam teori permainan, protokol komunikasi untuk melacak informasi turunan orde nol, orde pertama, dan orde kedua. Semantik fungsi dan representasi pengoptimalan diformalkan dalam tata bahasa untuk permainan guna menentukan bahasa formal untuk struktur algoritme pembelajaran mendalam. Model grafis probabilistik dan grafik faktor digunakan untuk menangkap fitur struktural distribusi multivariat serta desain dan analisisnya dalam algoritma untuk inferensi probabilistik.

Pengoptimalan berbasis gradien terutama ditujukan untuk memperoleh komputasi primitif yang dapat mengalihkan fokus desain arsitektur dari desain algoritme tingkat rendah di jaringan saraf ke desain mekanisme tingkat tinggi dalam sistem pembelajaran. Model komputasi kotak hitam untuk menganalisis kompleksitas komputasi metode optimasi adalah pandangan optimasi yang lebih abstrak daripada model mesin Turing. Ini menentukan protokol komunikasi yang melacak pola frekuensi tentang kueri yang dibuat oleh suatu algoritma untuk tujuan optimasi adversarial/pembelajaran. Di sini, metode optimasi kotak hitam orde nol merespons dengan informasi tentang nilai fungsi pada titik kueri, sedangkan metode optimasi kotak abu-abu orde pertama merespons dengan informasi tentang informasi gradien fungsi.

Rumusan teori permainan untuk tujuan pengoptimalan memungkinkan kerugian non-cembung yang dapat dirumuskan sebagai permainan pada berbagai skala berbeda dalam arsitektur jaringan saraf di mana lapisan tertentu dari jaringan saraf memecahkan masalah pengoptimalan cembung. Penataan aturan untuk komposisi tujuan pembelajaran mendalam yang bermusuhan kemudian dapat diformalkan dengan protokol komunikasi dan tata bahasa terdistribusi. Komputasi feedforward yang dihasilkan ditangkap dalam struktur data grafik komputasi yang menyusun kueri dan respons masing-masing menjadi grafik kueri dan respons. Protokol komunikasi menentukan bagaimana informasi penambangan data mengalir melalui grafik kueri dan respons tanpa menentukan kegunaan informasi bagi pemain.

Tata bahasa pada protokol komunikasi terdistribusi menjamin bahwa grafik respons mengkodekan informasi yang cukup bagi para pemain untuk bersama-sama menyatu ke konsep solusi teoretis permainan untuk fungsi tujuan pembelajaran dan fungsi hasil permusuhan yang terkait. Tata bahasa dapat ditentukan untuk interaksi pemain dan propagasi balik kesalahan di setiap game untuk melakukan tugas analisis data tertentu. Para pemain kemudian bersama-sama mengkodekan pengetahuan data mining tentang tugas tersebut. Tata bahasanya juga dapat mencakup formulasi probabilistik dan Bayesian serta

metode pra-pelatihan tanpa pengawasan. Contoh praktis tata bahasa ditunjukkan untuk tujuan pembelajaran dalam model pembelajaran mendalam yang diawasi, autoencoder variasional dan jaringan permusuhan generatif untuk pembelajaran tanpa pengawasan, dan model deviator-aktor-kritis untuk pembelajaran penguatan mendalam.

Hazan dkk. membahas minimalisasi penyesalan dalam permainan berulang dengan fungsi kerugian non-cembung. Permainan berulang seperti itu dapat digunakan untuk merancang permainan multipemain dalam pembelajaran mendalam yang bermusuhan. Gagasan tentang penyesalan dalam pembelajaran mendalam yang bermusuhan secara umum sulit dilakukan secara komputasi. Dengan demikian, formulasi penyesalan didefinisikan untuk optimasi yang efisien dan konvergensi ke perkiraan optimal lokal. Minimalkan penyesalan dalam permainan berhubungan dengan permainan berulang di mana pemain mengakumulasi kerugian rata-rata yang sebanding dengan keputusan respons terbaik jika dipikir-pikir. Penyesalan adalah kriteria pengoptimalan global yang dipilih oleh pemain di seluruh rangkaian keputusannya.

Jika fungsi kerugian yang menghitung imbalan pemain terhadap tindakan pemain lain adalah cembung, maka kriteria penyesalan tidak dapat diselesaikan secara komputasi, dan kriteria tersebut menyatu dengan konsep solusi teoretis permainan seperti ekuilibrium Nash, ekuilibrium berkorelasi, dan ekuilibrium berkorelasi kasar. Kriteria penyesalan lokal ditentukan untuk memprediksi poin permainan dengan rata-rata gradien kecil. Algoritme yang menimbulkan penyesalan lokal sublinear dalam waktu memiliki gradien rata-rata waktu yang kecil dalam ekspektasi untuk setiap iterasi yang dipilih secara acak. Gagasan penghalusan waktu menangkap pengoptimalan online non-cembung dalam penyimpangan konsep terbatas.

Sebaliknya, algoritme pengoptimalan kontinu non-cembung pada fungsi kerugian pemain fokus pada pencarian optimal lokal karena menemukan optimal global adalah masalah NP-hard. Metode stokastik orde kedua digunakan untuk optimasi non-cembung tersebut. Mereka menyatu pada konsep solusi yang kurang lebih stasioner dalam prosedur pembelajaran mendalam yang bermusuhan. Keseimbangan lokal diperhalus dibandingkan dengan iterasi sebelumnya. Prosedur pemulusan sesuai dengan bentuk pengulangan pengalaman dalam pembelajaran penguatan. Konsep solusi menangkap keadaan permainan yang berulang di mana setiap pemain memeriksa tindakan masa lalu yang dimainkan dan tidak ada pemain yang dapat melakukan penyimpangan untuk meningkatkan kinerja rata-rata permainan saat ini dibandingkan permainan historis lawannya. Algoritme pembelajaran diasumsikan memiliki akses ke oracle gradien stokastik yang berisik.

Xu dkk. menyelidiki masalah pembelajaran permusuhan dari data yang disuntikkan kebisingan tanpa mengasumsikan tipe musuh tertentu pada tahap pembelajaran. Batasan teoritis informasi dari ketahanan permusuhan yang disebut batas tipe Le Cam diturunkan. Karya ini sebanding dengan karya teoretis lainnya dalam teori pembelajaran komputasi untuk pembelajaran permusuhan seperti memperoleh batasan generalisasi untuk pembelajaran permusuhan pada waktu ujian, sertifikasi ketahanan untuk inferensi statistik dalam

pembelajaran permusuhan, kemampuan pembelajaran PAC yang kuat di kelas VC, dan analisis injeksi kebisingan di kelas VC.

pelatihan jaringan saraf pada waktu inferensi. Musuh diasumsikan memiliki anggaran berapa banyak noise yang dimasukkan ke dalam data. Anggaran ini terkait dengan total variasi (TV) jarak antara sebaran data asli dan sebaran data yang diinjeksi noise. TV adalah jarak statistik yang digunakan dalam studi batas atas dan bawah untuk ketahanan permusuhan dalam masalah pembelajaran seperti estimasi mean, klasifikasi biner, dan analisis Procrustes. Metode injeksi kebisingan dibatasi pada kebisingan multivariat Gaussian dan seragam multivariat. Risiko yang diharapkan diperkirakan dalam kerangka optimasi minimax untuk mendapatkan batasan Le Cam.

Scutari dkk. metode optimasi survei dalam sistem komunikasi dan pemrosesan sinyal. Model keseimbangan dalam teori permainan kooperatif dan non-kooperatif digunakan untuk menggambarkan skenario dengan keputusan interaktif dalam aplikasi seperti masalah komunikasi dan jaringan, kontrol daya dan pembagian sumber daya dalam jaringan nirkabel/kabel dan peer-to-peer, sistem radio kognitif, perutean terdistribusi , pengendalian aliran, dan kemacetan dalam jaringan komunikasi.

Teori ketidaksetaraan variasi (VI) digunakan sebagai kelas masalah umum dalam analisis non-linier aplikasi seperti itu dalam jaringan kabel ad hoc nirkabel atau per-to-peer, jaringan radio kognitif (CR), dan jaringan komunikasi multihop. Kemudian keberadaan dan keunikan keseimbangan teoritis permainan diselidiki untuk merancang sifat konvergensi dari algoritma terdistribusi berulang. Algoritme semacam itu juga dapat dirancang untuk pembelajaran mendalam permusuhan teoritis permainan dengan musuh variasional.

Hinrichs dkk. membahas pembelajaran transfer antar domain permainan sehingga analogi struktural dari satu permainan yang dipelajari mempercepat pembelajaran permainan terkait lainnya. Kenaikan minimal dan metamapping adalah teknik yang diusulkan untuk mentransfer representasi pencocokan analogi antara permainan dengan kosakata relasional yang berbeda. Kenaikan minimal menemukan hipotesis kecocokan lokal dengan memanfaatkan hubungan hierarki antar predikat. Metamapping adalah generalisasi kenaikan minimal untuk menggunakan semua informasi struktural yang tersedia tentang predikat dalam basis pengetahuan.

Domain permainan berkisar dari pemecahan masalah fisika hingga permainan strategi. Pembelajaran transfer didefinisikan sebagai masalah menemukan representasi analogi yang baik antara domain sumber dan target dan menggunakan representasi pengetahuan tersebut untuk menerjemahkan representasi simbolis dari pengetahuan yang dipelajari dari sumber ke target yang memiliki representasi permukaan yang sangat berbeda. Sebuah teori kognitif yang disebut pemetaan struktur digunakan untuk menggambarkan analogi yang mengikuti pemrosesan analogis manusia dan penilaian kesamaan.

Pemetaan tersebut mencakup inferensi kandidat yang mewakili informasi yang diproyeksikan dari domain sumber ke domain target. Kecocokan non-identik antar analogi dipertimbangkan ketika analogi tersebut merupakan bagian dari struktur relasional yang lebih besar yang dapat ditransfer. Jadi pemetaan struktur bergantung pada representasi data yang

simbolis dan terstruktur yang mencakup kosakata untuk merepresentasikan hierarki predikat, relasi himpunan, dan batasan pada tipe argumen terhadap predikat.

Predikat tingkat tinggi seperti penghubung logis, struktur argumen, perencanaan, dan hubungan wacana diasumsikan sama di seluruh domain sumber dan target. Konsep solusi teoretis permainan kemudian mendukung penalaran kualitatif dan analogis pada pemetaan dengan strategi komposisi dalam eksperimen transfer pada mesin keadaan terbatas. Mesin negara terbatas kemudian menciptakan pemahaman deklaratif tentang transfer pelajar yang paling efisien pada tingkat tindakan dan efek, ancaman dan bahaya, kemajuan menuju tujuan pembelajaran, dan analisis dinamis jejak permainan. Analisis domain statis dengan permainan mengarah pada perencanaan jalur dan perencanaan kuantitas dalam koordinat spasial, hubungan ordinal, operator pergerakan, dan potensi pengaruh pada kuantitas tersebut.

Analisis domain statis menghasilkan strategi komposisi ketika domain sumber-target bukan isomorf. Hal ini secara empiris membatasi ruang pencarian untuk strategi pembelajaran otomatis dalam heuristik keterjangkauan grafik. Sebaliknya, analisis domain dinamis memverifikasi asal usul strategi yang ditransfer untuk menggantikan strategi yang gagal dalam domain baru dengan tujuan pembelajaran. Analisis domain dinamis dengan permainan mengarah pada tujuan pembelajaran eksplisit untuk perolehan pengetahuan tentang dampak suatu tindakan, kondisi penerapan suatu tindakan, dan penguraian suatu tujuan menjadi sub-tujuan. Analisis domain dinamis beroperasi pada ruang pencarian tingkat yang lebih tinggi daripada representasi mesin negara untuk mendorong eksplorasi yang lebih efisien.

Analisis domain dinamis melakukan regresi melalui jejak eksekusi game untuk menjelaskan efek dan menyusun rencana untuk mencapai efek sesuai dengan heuristik preferensi. Urutan yang dipelajari dapat mengakomodasi dampak tindakan yang tidak terduga akibat respons permusuhan di mana perilaku musuh tidak diketahui secara lengkap. Eksperimen awal kemudian dapat dilakukan dari bawah ke atas untuk mempelajari efek tindakan dan prasyarat tindakan. Kemudian mereka dapat memasukkan tujuan pembelajaran tingkat tindakan untuk menguraikan tujuan kinerja permainan untuk mengembangkan strategi kemenangan dan pemberian kredit dengan penalaran otomatis. Perbaikan yang dilakukan melalui transfer ditandai dengan skor penyesalan yang dinormalisasi. Skor penyesalan yang lebih tinggi menunjukkan bahwa transfer tersebut bermanfaat.

Kemampuan untuk mempelajari lanskap pengoptimalan non-cembung adalah topik untuk pekerjaan masa depan dalam pembelajaran mendalam adversarial teoretis game. Kita dapat mengeksplorasi kriteria konvergensi dalam kesalahan generalisasi dan kompleksitas pengambilan sampel kelas konsep dalam pembelajaran mendalam adversarial teori permainan. Kita dapat menganalisis ukuran optimal data pelatihan untuk memprediksi perilaku masa depan dari fungsi target yang tidak diketahui dalam pembelajaran mendalam permusuhan sehingga fungsi hipotesis dalam optimasi teoretis permainan mungkin kira-kira benar. Di sini, penelitian kami terhadap model generatif mendalam dan autoencoder adversarial dapat diperluas menjadi studi tentang fungsi hipotesis dalam pembelajaran adversarial generatif.

Generalisasi keamanan siber dari pembelajaran permusuhan generatif mencakup struktur pemodelan dalam model generatif mendalam konvolusional, kondisional, dua arah, dan semi-supervisi seperti GAN dan autoencoder permusuhan. Menafsirkan kelas masalah optimasi stokastik kami sebagai bobot sinaptik dari variabel fuzzy mengarahkan kami untuk memperbarui aturan berdasarkan algoritma pembelajaran fuzzy yang menciptakan memori asosiatif di kelas konsep. Di sini, kami akan membandingkan solusi kami dengan dasar pembelajaran mesin seperti penyertaan noise dalam prosedur pengoptimalan, penyederhanaan lanskap fungsi dengan peningkatan ukuran model, skema pengoptimalan stokastik bebas turunan, dan pengambilan sampel ulang data dalam konteksnya. algoritma pembelajaran variasional.

Algoritma Pembelajaran Mendalam Manipulasi Game Theoretical dalam Aplikasi Perang Informasi

Pawlick dkk. mensurvei teori permainan untuk memodelkan penipuan defensif untuk keamanan siber dan privasi dalam komputasi yang ada di mana-mana dan dapat dipakai. Taksonomi penipuan diberikan sebagai gangguan, pergerakan pertahanan target, kebingungan, pencampuran, honey-x, dan keterlibatan penyerang. Ini mengkategorikan struktur informasi, agen, tindakan, dan durasi penipuan untuk pemodelan teori permainannya. Penelitian penipuan dilakukan dalam aplikasi militer, psikologi, kriminologi, keamanan siber, pasar ekonomi, advokasi privasi, dan ilmu perilaku.

Penipuan seperti ini biasa terjadi dalam interaksi keamanan siber yang bersifat permusuhan atau strategis, di mana satu pihak mempunyai informasi yang tidak diketahui pihak lain. Vektor serangan dengan penipuan seperti itu berpotensi mengubah perangkat Internet of Things menjadi senjata siber dalam negeri. Serangan siber dapat dirancang untuk secara fisik memengaruhi infrastruktur penting seperti jaringan listrik, mesin sentrifugal nuklir, dan bendungan air. Musuh memperoleh informasi tentang target mereka melalui pengintaian di mana penipuan melawan asimetri informasi. Teori permainan memodelkan interaksi yang menipu sebagai konfrontasi strategis atas konflik dan kerja sama antara agen-agen rasional. Setiap pemain dalam permainan keamanan siber dan privasi membuat keputusan yang memengaruhi kesejahteraan pemain lainnya.

Teori permainan mampu memodelkan aspek-aspek penting, dapat ditransfer, dan universal dari penipuan defensif di dunia maya. Permainan satu kesempatan dan banyak interaksi masing-masing mengarah pada penipuan statis dan dinamis. Teknik penipuan meliputi peniruan identitas, penundaan, pemalsuan, kamuflase, alasan palsu, dan rekayasa sosial. Tahapan dalam penipuan jahat meliputi desain cerita sampul, perencanaan, pelaksanaan, dan pemantauan. Stackelberg, Nash, dan permainan sinyal adalah model teoritis permainan yang paling umum dengan interaksi dinamis dua pemain. Domain aplikasi mencakup pembelajaran mesin permusuhan, sistem deteksi intrusi, gangguan komunikasi, dan keamanan bandara.

Nguyen dkk. menyelidiki penipuan strategis dari musuh yang memiliki informasi pribadi dalam berinteraksi berulang kali dengan pembela HAM. Informasi pribadi sensitif yang dimiliki musuh mewakili domain keamanan dunia nyata yang ditandai dengan informasi yang

tidak sempurna karena ketidakpastian mengenai tindakan dan karakteristik lawan. Di sini, perilaku menipu dapat memanipulasi hasil pembelajaran demi keuntungan jangka panjang dari musuh yang manipulatif. Penerapan strategi penipuan permusuhan dapat dimodelkan dengan permainan keamanan yang diulang-ulang secara terbatas.

Pembela mempunyai informasi yang tidak lengkap dan ketidakpastian mendasar mengenai tipe musuh. Pada setiap iterasi interaksi teoritis permainan antara musuh dan pembela, pembela memperbarui keyakinan tentang tipe musuh berdasarkan data serangan historis yang dikumpulkan pada iterasi sebelumnya. Bek memilih tindakan untuk dimainkan berdasarkan keyakinan yang diperbarui. Tujuan musuh adalah untuk mengoptimalkan utilitas yang diharapkan yang mengarah pada keseimbangan Bayesian Nash yang sempurna.

Ferguson-Walter dkk. membahas bagaimana penipuan dalam pertahanan dunia maya menyeimbangkan kerugian asimetris dengan mempersulit pekerjaan musuh. Penipuan dunia maya menambah ketidakpastian tentang informasi yang sebenarnya dengan menambahkan informasi yang salah. Penipuan dunia maya berdampak pada pengambilan keputusan pihak lawan sehingga menyia-nyiakan waktu, tenaga, dan sumber daya mereka. Penipuan dunia maya yang digunakan oleh pembela HAM memberikan keyakinan yang salah kepada musuh di setiap tahap rantai pembunuhan dunia maya dalam serangan multi-tahap multi-vektor.

Kecerdasan buatan, keamanan komputer, dan ilmu perilaku dalam pertahanan siber adaptif atau aktif secara proaktif dan dinamis menerapkan strategi pertahanan prediktif tanpa campur tangan manusia. Kejutan karena hasil yang tidak terduga merupakan elemen penting untuk mengganggu atau menunda proses pengambilan keputusan dan tindakan penyerang sehingga memberikan lebih banyak waktu dan kesempatan bagi pembela untuk merespons dan bereaksi. Teknik adaptif kemudian mendeteksi respons penyerang terhadap penipuan dunia maya, untuk mengubah metode penipuan tersebut. Selain kejutan, menyebabkan frustrasi, kebingungan, dan keraguan diri adalah cara lain untuk mempengaruhi penyerang dalam penipuan dunia maya yang mengeksploitasi bias kognitif musuh untuk menciptakan beban kognitif yang berlebihan.

Dalam penelitian penipuan dunia maya, sistem umpan mencakup honeypots dan honey-token, serangan replay, pembuatan paket dan perubahan muatan, tar-pitting, dan dokumen palsu. Mereka menciptakan lingkungan umpan bagi musuh dengan sistem virtual yang realistis dan ringan yang tampak seperti sistem nyata yang menjalankan layanan nyata. Mereka dikerahkan bersamaan dengan sistem nyata untuk meningkatkan peluang musuh terdeteksi dan dimitigasi dengan cepat. Mereka juga memberikan keuntungan asimetris kepada para pembela siber dengan mengurangi kemungkinan aset nyata diserang oleh musuh yang teralihkan perhatiannya. Mereka juga meningkatkan peluang musuh untuk mengungkapkan diri mereka dengan mengambil tindakan tambahan.

Sistem penipuan siber otonom menyediakan penipuan siber yang adaptif terhadap strategi dan preferensi masing-masing jenis musuh dalam lanskap pertahanan siber. Strategi penipuan dunia maya memerlukan sensor dan aktuator yang mengambil keputusan tentang bagaimana dan kapan harus beradaptasi. Sensor mengumpulkan aktivitas permusuhan pasca-eksploitasi berbasis perilaku seperti aktivitas pemindaian, upaya login, dan kata sandi yang

dicuri, serta menyebarkan umpan seperti token madu. Aktuator mengambil tindakan otomatis pada jaringan atau host. Aktuator umpan membuat perubahan konfigurasi, mengubah alamat IP, membuka/menutup port, menambah/menghapus layanan, memalsukan sistem operasi, dan membuat umpan baru.

Sistem pertahanan siber adaptif tersebut harus mempertimbangkan evolusi bersama dari situasi serangan/pertahanan multi-tahap dan multi-tahap di mana gerakan pembela disimulasikan beberapa langkah sebelum tindakan penyerang. Di sini, sasaran bek tingkat lanjut mencakup perolehan preferensi tentang sasaran penyerang dan kesalahan informasi topologi tentang keyakinan yang salah dalam topologi jaringan. Mereka dapat dimasukkan ke dalam tugas klasifikasi permusuhan dengan penguatan dalam pembelajaran mendalam permusuhan teoritis permainan dengan hypergames di mana algoritma pembelajaran memprediksi preferensi serangan.

Dalam teori permainan, utilitas penipuan dan kesalahan persepsi yang disengaja diformalkan sebagai hypergames di mana penipuan adalah komponen strategi yang dimainkan. Hypergames merumuskan tujuan pembela, pengamatan, subpermainan, dan strategi individu yang ditentukan dalam konteks permainan yang terdiri dari konteks musuh dan konteks pembela yang menyajikan perspektif permainan yang spesifik untuk pemain. Melalui observasi terhadap penyerang, pembela HAM mencoba menyimpulkan keyakinan penyerang dari waktu ke waktu dan menerapkannya dalam pengambilan keputusan di masa depan. Keyakinan penyerang digunakan untuk memperkirakan keadaan tipe musuh serta imbalan yang dirasakan penyerang dengan pengetahuan tentang pohon permainan dan persepsi penyerang.

Pembela kemudian secara dinamis memanipulasi papan permainan dengan aturan pembaruan untuk mengubah hasil berulang yang terkait dengan tindakan selanjutnya yang mungkin dilakukan. Keputusan yang dibuat oleh pembela HAM dalam solusi pembelajaran online mengubah tindakan yang diambil oleh penyerang, membatasi strategi yang tersedia bagi penyerang pada langkah berikutnya, dan memanipulasi imbalan yang diterima oleh penyerang. Dengan demikian, konsep hypergame dapat menyelidiki pohon serangan berdasarkan tujuan pembela pada penipuan daripada tujuan permusuhan pada manipulasi di mana terdapat biaya alokasi sumber daya yang terkait dengan setiap permainan.

Cybenko dkk. mengedit buku tentang teknik adaptasi (AT) seperti memindahkan pertahanan target (MTD) untuk merekayasa sistem pembelajaran mesin permusuhan dengan pengacakan untuk tujuan keamanan dan ketahanan. Pertahanan cyber adaptif (ACD) dikategorikan ke dalam teknik adaptasi (AT) dan penalaran adversarial (AR) untuk pembelajaran adversarial dalam sistem pembelajaran operasional. AR menggabungkan pembelajaran mesin, ilmu perilaku, riset operasi, teori kontrol, dan teori permainan untuk menghitung strategi dalam lingkungan yang dinamis dan bermusuhan. Teknik ACD memaksa musuh untuk menilai ulang, merekayasa ulang, dan meluncurkan kembali serangan siber. Analisis permainan teoretis dan teori kontrol untuk analisis trade-off persyaratan keamanan di ACD menghadirkan permukaan serangan yang dioptimalkan dan berubah secara dinamis

kepada musuh. Prototipe dan demonstrasi teknologi ACD disajikan dalam beberapa skenario dunia nyata.

Dasgupta dkk. melakukan survei berkualitas tinggi tentang pemodelan pembelajaran permusuhan berbasis teori permainan. Mekanisme prediksi algoritma pembelajaran mesin yang diawasi dirangkum. Namun ide tersebut dapat diterapkan pada mekanisme pembelajaran mesin lainnya dalam pengelompokan, pemeringkatan, atau regresi. Taksonomi untuk pembelajaran mesin permusuhan dicirikan berdasarkan pengaruh, kekhususan, dan dimensi pelanggaran keamanan di seluruh jenis musuh. Dimensi pengaruh menentukan serangan kausatif dan eksplorasi terhadap kerentanan pelajar untuk membuat data pelatihan yang dimodifikasi dan data pengujian yang disebut data adversarial. Menambahkan data permusuhan ke proses pembelajaran klasifikasi akan menyebabkan pengklasifikasi yang salah sehingga menghasilkan kesalahan klasifikasi.

Pembelajaran permusuhan untuk membuat pengklasifikasi yang aman kemudian dimodelkan sebagai permainan dua pemain dan non-kooperatif. Fungsi utilitas setiap pemain mengungkapkan preferensi pemain terhadap berbagai hasil permainan yang dinyatakan dalam tindakan bersama semua pemain dalam permainan. Hasil dari suatu permainan adalah strategi yang dipilih oleh setiap pemain. Kriteria optimasi yang paling populer untuk menghitung hasil adalah ekuilibrium Nash yang mengasumsikan hasil merupakan strategi respons terbaik dari pemain rasional. Ekuilibrium Nash diselesaikan sebagai masalah pencarian dan optimasi. Dalam permainan dua pemain, zero-sum, utilitas semua pemain berjumlah nol pada setiap iterasi permainan. Dalam pembelajaran adversarial dengan permainan zero-sum, perolehan utilitas bagi pembelajar mengakibatkan hilangnya utilitas musuh dan sebaliknya. Pengamatan ini mengarah pada teorema minimax untuk mencari keseimbangan Nash dalam permainan zero-sum. Hasil minimax direpresentasikan sebagai masalah optimasi terbatas yang diselesaikan dengan program linier.

Teorema minimax tidak berlaku untuk permainan jumlah umum bukan jumlah nol. Karena pengklasifikasi bereaksi terhadap manipulasi permusuhan, pemilihan strategi dalam pembelajaran permusuhan paling sering dimodelkan sebagai permainan gerak berurutan daripada permainan gerak simultan. Dalam permainan gerak sekuensial, pemain pengikut mempunyai informasi tentang strategi yang dipilih oleh pemain pemimpin. Informasi tersebut digunakan dalam optimalisasi fungsi utilitas pemain. Namun, pemimpin harus memperhitungkan ketidakpastian tentang strategi pengikutnya yang mengarah pada permainan Bayesian. Dalam permainan Bayesian bentuk normal, setiap pemain memiliki informasi tentang utilitas pemain pesaing lainnya. Berdasarkan informasi ini, kita dapat menghitung utilitas yang diharapkan berdasarkan tipe pemain untuk setiap pemain.

Permainan keamanan dalam keamanan siber terkait dengan permainan Bayesian untuk pembelajaran permusuhan. Dalam permainan keamanan, pelajar adalah pembela yang melindungi serangkaian target dari musuh yang disebut penyerang. Pembela HAM harus melakukan alokasi sumber daya sesuai batasan anggaran dan operasional. Secara umum, utilitas pembelajar dihitung berdasarkan hasil/nilai pembelajar dalam mengklasifikasikan input dengan benar. Demikian pula, utilitas musuh didefinisikan sebagai imbalan/nilai

kesalahan klasifikasi masukan musuh yang diberikan kepada pelajar. Masalah pembelajaran tersebut kemudian dirumuskan sebagai masalah optimasi terbatas. Ini diselesaikan sebagai program linier bilangan bulat campuran untuk musuh dan strategi klasifikasi yang kuat untuk pelajar.

Pendekatan adversarial yang menghindari rekayasa balik batasan keputusan pengklasifikasi kotak hitam menelusuri ruang biaya adversarial untuk menentukan serangkaian contoh adversarial minimum. Memindahkan pertahanan target akan memperluas pembelajar yang dihasilkan untuk menggunakan pengacakan pada beberapa pengklasifikasi alih-alih menyetel parameter dari satu pengklasifikasi yang kuat. Musuh juga dapat menghasilkan data dengan memilih, menghapus, atau merusak fitur dari kumpulan data masukan. Tujuan pembelajar adalah menemukan serangkaian fitur optimal yang meminimalkan fungsi kerugiannya. Ketika pembelajar tidak memiliki akses terhadap seluruh data pelatihan, maka tujuan pembelajaran menjadi masalah pembelajaran online.

Kami kemudian dapat menganalisis runtime dan kompleksitas sampel pembelajar online yang diadu dengan tipe musuh yang berbeda dengan fungsi biaya adversarial dan fungsi pembuatan contoh adversarialnya sendiri. Di sini, pelajar dapat mengetahui tentang fungsi biaya adversarial tetapi tidak mengetahui kebenaran dasar atau distribusi inputnya. Pelatihan permusuhan jaringan saraf dalam mengarah pada permainan pembelajaran mendalam. Hal ini dapat dirumuskan sebagai permainan zero-sum yang berulang.

Iterasi pada pemutaran berulang kemudian digunakan untuk menyesuaikan bobot tepi jaringan saraf agar menyatu ke keseimbangan Nash dengan algoritma pengoptimalan pencocokan bobot dan penyesalan yang dieksponensiasi. Ketahanan adversarial dari pengklasifikasi pembelajaran mendalam juga dapat ditingkatkan dengan generator data adversarial yang digunakan bersama dengan prosedur pembelajaran adversarial. Generator data adversarial yang paling umum menggunakan teknik perturbasi pada contoh yang valid, mentransfer contoh adversarial ke berbagai model pembelajar, dan memperluas jaringan adversarial generatif.

Di sini, model teori permainan dapat dirumuskan untuk diinformasikan melalui pemodelan dan penalaran biaya seperti biaya untuk menyelesaikan keseimbangan Nash, biaya untuk mempertahankan sejarah permainan, biaya untuk membangun model lawan dari sejarah interaksi teoritis permainan, dan biaya yang dikeluarkan oleh musuh untuk mengakses sumber daya yang sah. Pembelajaran transfer dapat dikombinasikan dengan pembelajaran adversarial dalam aplikasi dunia nyata untuk menciptakan sistem pembelajaran pada data pelatihan renggang yang membuat prediksi klasifikasi dengan benar tanpa memerlukan sumber data yang kaya informasi. Adaptasi domain dapat diterapkan pada pembelajaran adversarial untuk mentransfer pembelajar kuat yang dipetakan dalam domain sumber padat ke domain target yang jarang secara andal. Aplikasi untuk menggabungkan pembelajaran transfer dengan pembelajaran permusuhan mencakup pengklasifikasi spam email, alat analisis sentimen jaringan sosial, dan sistem pengenalan data gambar dan sensor pada kendaraan otonom.

Hamilton dkk. membahas penerapannya dalam analisis taktis peperangan informasi. Algoritme teori permainan dapat dikembangkan dalam aplikasi militer untuk memprediksi serangan di masa depan dalam berbagai kemungkinan skenario dan menyarankan tindakan (COA) sebagai respons terhadap kemungkinan yang paling berbahaya. Teknik pembangkitan COA bisa mendapatkan keuntungan dari pembelajaran permusuhan. Kerangka teoritis permainan memungkinkan analisis rinci tentang skenario bagaimana-jika rangkaian peristiwa untuk menemukan pengecualian terhadap aturan umum dalam sistem permainan dunia maya.

Analisis tersebut menentukan kemungkinan, metode, dan biaya skenario seperti pengumpulan intelijen dalam fase serangan, penargetan sistem komando dan kontrol, korupsi data, dan serangan penolakan layanan untuk mencegah proses perencanaan peperangan kinetik. Teknik pemangkasan diperlukan untuk mengurangi ruang pencarian dalam mengevaluasi permainan max-max yang kompleks sehingga node yang paling menjanjikan dalam pohon permainan diperluas untuk mengantri anak-anaknya untuk analisis langkah paling menjanjikan yang paling mungkin diprediksi dalam suatu waktu tertentu. skenario dunia nyata.

Teknik pembelajaran penguatan dapat digunakan untuk meningkatkan kedalaman pohon permainan secara berulang dan menetapkan karakteristik evaluasi yang sesuai untuk mempelajari kedalaman mana yang paling baik dalam memprediksi perilaku lawan. Di sini asumsi dalam perancangan fungsi evaluasi pembela adalah bahwa fungsi evaluasi lawan menggunakan subset heuristik dari fungsi evaluasi pembela dengan perubahan bobot optimalitas. Schlenker dkk. memperkenalkan permainan penipuan dunia maya dalam keamanan jaringan. Hal ini diselesaikan dengan solusi program linier bilangan bulat campuran dan algoritma pencarian minimax serakah yang cepat.

Kerangka pembelajaran permusuhan teoretis permainan untuk kebingungan layanan ini dapat digunakan secara dinamis untuk membuat informasi asimetris tentang keadaan sebenarnya dari jaringan selain tindakan statis keamanan jaringan seperti memasukkan aplikasi ke daftar putih, mengunci izin, dan menambal kerentanan. Pekerjaan terkait adalah dalam permainan pemilihan honeypot, permainan sinyal musuh, dan model logika probabilistik yang dianotasi pada permintaan pemindaian penyerang. Permainan penipuan dunia maya adalah permainan Stackelberg zero-sum antara pembela administrator jaringan dan musuh peretas. Nguyen dkk. mengalokasikan tindakan pencegahan keamanan terbatas untuk melindungi data jaringan dari skenario serangan siber yang dimodelkan sebagai grafik serangan Bayesian.

Interaksi multi-tahap antara administrator jaringan dan penjahat dunia maya dirumuskan sebagai permainan keamanan. Strategi heuristik berparameter dihitung untuk mengeksploitasi struktur topologi grafik serangan. Metode pengambilan sampel digunakan untuk mengatasi kesulitan kompleksitas komputasi dalam memprediksi tindakan lawan. Dalam keamanan komputer, grafik serangan adalah model grafis yang menguraikan ontologi skenario keamanan kompleks menjadi taksonomi tindakan sederhana dan terukur. Grafik serangan pada tindakan musuh cocok untuk merancang pertahanan target bergerak (MTD)

bagi pemain bertahan di mana taktik proaktif digunakan untuk mengubah konfigurasi sistem secara dinamis. Formalisme grafik serangan Bayesian memodelkan masalah pembelajaran mendalam yang bermusuhan sebagai permainan keamanan grafik serangan multi-tahap secara simultan. Titik-titik pada grafik serangan mewakili kondisi keamanan sistem jaringan. Tepi mewakili hubungan antar kondisi keamanan.

Grafik serangan mewakili kerentanan jaringan dalam skenario keamanan yang kompleks. Pembela melindungi sekumpulan node sasaran dalam grafik serangan dengan tindakan pencegahan keamanan yang diterapkan. Sementara itu, kemajuan penyerang memilih simpul dalam grafik serangan. Masalah keamanan dianalisis sebagai solusi keseimbangan permainan dinamis dan permainan stokastik pada grafik serangan dengan informasi lengkap/tidak lengkap/tidak sempurna yang dapat diamati sebagian. Heuristik berparameter memperkirakan nilai serangan untuk setiap simpul berdasarkan nilai serangan dari simpul tetangga, di mana nilai serangan sesuai dengan pentingnya setiap simpul dalam permainan keamanan. Jalur serangan untuk penyerang dipilih dengan metode sampling.

Pada setiap langkah waktu, strategi pertahanan diperbarui dengan penyaringan partikel untuk mencerminkan keyakinan pemain bertahan tentang hasil tindakan pemain pada langkah waktu sebelumnya. Tindakan pertahanan baru dihasilkan dari konsep solusi permainan keamanan berdasarkan keyakinan terkini dan asumsi pembela HAM tentang strategi penyerang. Kekuatan strategi pertahanan dalam menghadapi musuh bergantung pada ketidakpastian mengenai kondisi permainan dan strategi penyerang. Permainan keamanan yang kompleks tersebut merangkum lingkungan keamanan dinamis dengan ketidakpastian sebagai proses stokastik dalam beberapa langkah waktu.

Kulkarni dkk. membahas masalah perencanaan dalam sistem AI. Rencana kebingungan dijalankan dalam situasi yang merugikan untuk melindungi privasi. Pengaturan permusuhan mencakup perencanaan misi, intelijen militer, pengintaian, dll. Rencana yang terbaca dilaksanakan dalam situasi kooperatif untuk membantu pemahaman. Rencana yang dikaburkan konsisten dengan setidaknya k sasaran dari serangkaian sasaran umpan di akhir rangkaian observasi. Rencana yang terbaca konsisten dengan paling banyak j sasaran dari serangkaian sasaran perancu di akhir rangkaian observasi.

Rencana dihitung dari sudut pandang pengamat yang memiliki informasi terbatas yang beroperasi dalam ruang kepercayaan. Untuk setiap tindakan yang diambil oleh agen pembelajaran dan transisi keadaan terkait, pengamat menerima observasi. Pekerjaan terkait adalah pelestarian privasi dalam sistem multi-agen terdistribusi, perencanaan gerak, dan robotika. Kebingungan tujuan terkait dengan literatur pengenalan rencana untuk observasi keadaan tindakan yang bising. Model “Perencanaan yang dapat dijelaskan” adalah pemahaman manusia yang mengamati atau berinteraksi tentang agen perencanaan untuk perencanaan multi-model yang disadari oleh manusia.

Zhang dkk. mensurvei titik temu antara privasi diferensial dan teori permainan untuk membangun manipulasi permusuhan dalam pembelajaran mesin dengan desain mekanisme. Manipulasi permusuhan tersebut mempertimbangkan trade-off biaya-manfaat antara pelanggaran privasi dan pelanggaran keamanan dalam merancang kekuatan permusuhan.

Privasi diferensial membatasi keuntungan yang diharapkan yang dapat diperoleh dari manipulasi strategis fungsi imbalan yang merugikan dalam permainan keamanan. Agen yang sadar privasi dalam permainan keamanan memiliki preferensi privasi yang dirumuskan secara eksplisit dalam fungsi imbalan mereka yang sadar privasi. Jadi mereka bisa membuat trade-off antara privasi dan utilitas.

Masalah desain mekanisme dengan privasi diferensial untuk agen yang sadar privasi dapat dikombinasikan dengan masalah optimasi dalam teori permainan. Hal tersebut mencakup pemilihan strategi, implikasi informasi, insentif kebenaran, estimasi biaya privasi, perdagangan data pribadi, dan pembelajaran permainan. Kamu dkk. mengusulkan pendekatan teoretis permainan pribadi yang berbeda terhadap penipuan dunia maya. Pembela menggunakan mekanisme privasi diferensial untuk mengaburkan konfigurasi sistem. Penyerang menggunakan inferensi Bayesian untuk menyimpulkan konfigurasi sistem yang sebenarnya. Permainan informasi yang tidak sempurna digunakan sebagai permainan penipuan dunia maya dengan tujuan menyembunyikan informasi tentang konfigurasi sistem daripada menghentikan serangan dunia maya atau mengidentifikasi penyerang. Hal ini dapat diperluas ke permainan keamanan dinamis dengan banyak pembela dan banyak penyerang.

Ras dkk. menyelidiki permainan matriks sebagai alat mitigasi risiko untuk pertahanan ancaman persisten tingkat lanjut (APT). APT menggabungkan beberapa vektor serangan seperti rekayasa sosial atau malware dari analisis kerentanan topologi untuk menghasilkan ketidakpastian penilaian risiko ahli kualitatif, insentif permusuhan yang tidak diketahui, dan keadaan sistem saat ini. Pembelajaran permusuhan teoritis permainan mengoptimalkan pertahanan simultan melawan penyerang diam-diam menggunakan serangkaian jalur yang diketahui dalam serangan APT. Jalur serangan adalah serangkaian kerentanan. Pengukuran entropi grafik pada jalur serangan yang terlewat dapat mengukur ketidakpastian kompleksitas jaringan dan risiko sisa dari eksploitasi kerentanan yang belum diketahui sehingga menghasilkan jalur serangan yang disebut eksploitasi zero-day.

Eksploitasi zero-day ditangani dengan kombinasi pengetahuan domain, pendapat ahli, pengalaman, dan penambangan informasi, dikombinasikan dengan model matematika yang sesuai dari ambang batas penerimaan risiko dan distribusi fungsi kerugian dalam pembelajaran mesin permusuhan. Kecerdasan ancaman dunia maya dan keahlian domain adalah konsep dasar yang menjadi dasar pengukuran distribusi fungsi kerugian secara empiris berdasarkan penalaran dan pengalaman manusia. Penalaran manusia seperti itu tidak jelas dan tidak jelas sehingga mengarah pada distribusi data multimodal. Ukuran kinerja numerik untuk pembelajaran mesin tidak cukup untuk menangani serangan APT karena banyak data statistik yang dapat diandalkan mengenai insiden keamanan siber tidak tersedia dalam aplikasi dunia nyata.

Penanggulangan terhadap APT dapat diidentifikasi namun tidak selalu mungkin, layak, atau berhasil. Jadi mekanisme pertahanan harus mengasumsikan perilaku terburuk penyerang yang mungkin tidak sesuai dengan formulasi permainan Bayesian atau sekuensial. Permainan mitigasi APT adalah waktu diskrit untuk pemain bertahan dan waktu berkelanjutan untuk penyerang. Data adversarial adalah pengetahuan ahli fuzzy kualitatif yang dirumuskan

dari taksonomi atau simulasi atau keduanya. Ada asimetri yang kuat antara informasi yang diketahui penyerang dan pembela. Permainan mitigasi APT seperti ini dimodelkan sebagai permainan informasi yang lengkap namun hasil yang tidak pasti.

Imbalan yang tidak pasti adalah distribusi probabilitas, bukan angka riil. Pemodelan permainan mitigasi APT dapat diperluas ke permainan multi-kriteria dengan trade-off optimal antara beberapa tujuan yang mencakup persyaratan keamanan seperti kerahasiaan vs. integritas vs. ketersediaan, fungsi biaya yang berlawanan dalam grafik serangan tentang tingkat keterampilan yang diperlukan untuk melakukan serangan, dan serangkaian kategori kerugian tetap yang berlaku untuk semua tujuan permusuhan dan pembelajaran yang diinginkan. Di sini, struktur permainan diasumsikan mengikuti proses stokastik sehingga distribusi kerugian yang membentuk struktur permainan adalah distribusi stasioner dari proses stokastik berdasarkan kriteria konvergensi yang dipilih.

Huang dkk. mengusulkan kerangka permainan dinamis untuk interaksi jangka panjang antara penyerang diam-diam dan pembela proaktif yang dirumuskan sebagai permainan multi-tahap informasi tidak lengkap di mana setiap pemain memiliki informasi pribadi yang tidak diketahui satu sama lain. Para pemain bertindak secara strategis sesuai dengan keyakinan yang dibentuk oleh observasi dan pembelajaran multi-tahap. Keseimbangan Bayesian Nash yang sempurna adalah konsep solusi yang dihitung dengan algoritma optimasi berulang. Penyerang siluman adalah penyerang APT yang memiliki pengetahuan tentang arsitektur sistem pembela, aset berharga, dan strategi pertahanan. Jadi strategi penyerang APT dibuat khusus untuk membatalkan kriptografi, firewall, dan sistem deteksi intrusi. Penyerang APT dapat menyamar sebagai pengguna sah dalam jangka panjang.

Model APT multi-tahap yang membagi rangkaian serangan diklasifikasikan ke dalam rangkaian serangan, atau fase tersedia di komunitas intelijen sumber terbuka. Hal ini mencakup rantai pembunuhan siber Lockheed-Martin, ATT&CK MITRE, dan kerangka kerja ancaman siber teknis NSA/CSS. Selama fase pengintaian, penyerang (juga disebut aktor ancaman) mengumpulkan intelijen sumber terbuka atau internal untuk mengidentifikasi target yang berharga. Kemudian penyerang meningkatkan hak istimewa untuk menyebar secara lateral di jaringan cyber untuk mengakses informasi rahasia atau menimbulkan kerusakan fisik. Seorang pembela sistem harus menggabungkan tindakan pencegahan defensif di seluruh fase APT dengan strategi pertahanan yang mendalam.

Dalam mengidentifikasi utilitas dan strategi penyerang, teori permainan memberikan kerangka kerja kuantitatif dan dapat dijelaskan kepada pembela sistem untuk merancang respons pertahanan proaktif dalam ketidakpastian dengan keseimbangan yang lebih baik antara keamanan dan kegunaan. Sebaliknya, metode pertahanan berbasis aturan dan pembelajaran mesin tidak dapat mengatasi ketidakpastian dampak multi-tahap strategi pertahanan terhadap pengguna yang sah dan pihak yang bermusuhan. Dampak multi-tahap seperti ini terlihat dalam kecerdasan buatan, ekonomi, dan ilmu sosial di mana interaksi multi-tahap terjadi antara banyak agen dengan informasi yang tidak lengkap.

Bohrer dkk. menerapkan logika permainan diferensial konstruktif untuk mendapatkan bukti terstruktur dalam sistem cyber-fisik (CPS) untuk aplikasi penting keselamatan seperti

robotika, otomotif, penerbangan, penerbangan luar angkasa, perangkat medis, dan sistem tenaga. Metode verifikasi formal tersebut memastikan kebenaran sifat pembelajaran model sistem dalam implementasi CPS pada prosesor tertanam seperti dalam mengemudi otonom dan robotika darat. Di sini, teori permainan digunakan dalam analisis persamaan diferensial tanpa solusi bentuk tertutup. Bukti permainan dalam lingkungan permusuhan kemudian menciptakan jaminan keamanan untuk sistem pembelajaran terhadap berbagai jenis musuh yang melanggar kriteria kebenaran sistem dengan manipulasi waktu, penginderaan, kontrol, dan fisika dalam permainan hibrida.

Wellman dkk. mengeksplorasi struktur ketergantungan kausal dalam pola sinyal informasi pribadi pada keadaan agen yang mendasarinya yang dapat bertindak sebagai jenis agen permusuhan yang epistemik. Permainan Bayesian kemudian dirumuskan dengan informasi pribadi. Model grafis probabilistik (PGM) digunakan untuk memodelkan informasi pribadi. Struktur ketergantungan mereka mampu mengkuantifikasi agen-agen yang bermusuhan yang mempertimbangkan nilai-nilai fungsi pembayaran mereka sendiri berdasarkan informasi yang tersedia mengenai nilai-nilai fungsi pembayaran dari agen-agen lain. Jaringan Bayesian adalah PGM yang digunakan tidak hanya untuk menafsirkan tetapi juga menghasilkan pola sinyal pada informasi pribadi pemain.

Dengan demikian, PGM memberikan kerangka kualitatif untuk menganalisis ketergantungan probabilistik antara keputusan struktural dan situasi pembelajaran teoritis permainan sinyal swasta. Struktur grafisnya memungkinkan penalaran tentang implikasi situasi permainan. Aplikasi dievaluasi untuk pasar prediksi dan lelang yang menambah struktur grafis dengan pola penalaran yang tersedia di domain ini untuk menyarankan generalisasi hasil dalam pembelajaran mendalam adversarial yang berlaku. Yordania dkk. melakukan analisis empiris terhadap permainan kompleks. Nilai empirisme dalam permainan terletak pada eksplorasi efektif terhadap serangkaian strategi. Jadi kebijakan eksplorasi generik diusulkan untuk eksplorasi strategi dalam permainan empiris.

Mereka menemukan respons terbaik dengan profil penyesalan minimal di antara strategi yang telah dieksplorasi sebelumnya. Strategi respons terbaik stokastik mengarah pada eksplorasi ruang strategi yang efektif. Jadi analisis teori permainan empiris (EGTA) dapat menambah pemodelan ahli dengan sumber pengetahuan empiris seperti data dengan ketelitian tinggi yang diperoleh dari observasi dunia nyata. Permainan EGTA adalah deskripsi prosedural dari lingkungan strategis. Statistik simulasi dan penelusuran di EGTA dapat digabungkan dengan konsep solusi teoretis permainan untuk mengkarakterisasi properti strategis domain aplikasi untuk pembelajaran mendalam yang bermusuhan. Game terbatas yang diperbesar didefinisikan sebagai game dasar untuk merangkum EGTA. Strategi tambahan dihasilkan dengan pembelajaran penguatan.

Prakash dkk. meneliti interaksi antara strategi serangan dan pertahanan dalam pertahanan target bergerak (MTD). Beberapa contoh permainan dieksplorasi berdasarkan perbedaan dalam tujuan agen, biaya serangan, dan detektor tindakan serangan. Teknik MTD tersebut menggabungkan perkembangan serangan probabilistik untuk mengembangkan kebijakan yang efektif dalam menerapkan dan mengoperasikan sistem pembelajaran mesin

dalam konteks permusuhan tertentu. Perilaku pemain rasional bervariasi menurut fitur pembelajaran teoretis permainan seperti konfigurasi sistem, kondisi lingkungan, tujuan agen, dan karakteristik teknologi.

Simulasi sistematis dalam kerangka pembelajaran teori permainan dapat mengakomodasi kompleksitas komputasi dan ketidakpastian informasi dalam dinamika pembelajaran formulasi permainan yang sulit diselesaikan secara analitis. Fungsi imbalan yang berlawanan dalam MTD memungkinkan terjadinya trade-off antara tujuan pengendalian dan ketersediaan. Mereka dapat menggabungkan penilaian keadaan sistem secara keseluruhan sebagai fungsi biaya yang merugikan. Selain tujuan pembelajaran musuh dan pembela, persyaratan keamanan sistem pembelajaran ditafsirkan sebagai pola preferensi agen dalam permainan MTD. Misalnya, kerahasiaan sistem pembelajaran ditafsirkan sebagai keengganan kuat pembela HAM untuk mengizinkan penyerang mengendalikan server pembelajaran mesin.

Availability diartikan sebagai kendali bek pada sebagian kecil server yang tidak down. Skema pembobotan kemudian memasukkan trade-off antara kerahasiaan dan ketersediaan dalam formulasi permainan. Kelompok strategi heuristik yang diparameterisasi ditentukan oleh struktur dan pola perilaku dari waktu ke waktu sehingga strategi tersebut menentukan kebijakan tindakan yang melaksanakan pilihan pemain di antara strategi. Permainan terbatas pada serangkaian strategi yang dipilih kemudian secara sistematis menyempurnakan eksplorasi strategi melalui proses analisis permainan empiris yang berulang. Profil strategi dan kriteria validasi kemudian muncul berdasarkan proses penalaran teoritis permainan dalam lingkungan permusuhan yang kompleks. Hal ini dapat diperluas ke arah kecanggihan dalam kebijakan penyerang dan pembela dengan inferensi yang disengaja, alasan eksplisit tentang ancaman dan kontra-ancaman, dan model utilitas untuk waktu henti stokastik.

Roeder dkk. mengusulkan metode untuk mengurangi kerentanan bersama antar server yang disebut kebingungan proaktif. Transformasi kode yang mempertahankan semantik digunakan untuk menghasilkan beragam executable yang membatasi jumlah server yang disusupi dengan memulai ulang secara berkala. Kebingungan proaktif membuat pekerjaan musuh menjadi lebih sulit dengan memulai ulang server secara acak ke kondisi baru. Hal ini ditunjukkan dalam firewall terdistribusi dan penyimpanan terdistribusi berdasarkan replikasi mesin negara. Biaya yang tersirat dalam kebingungan proaktif dievaluasi dengan mengukur kinerja sistem. Teknik kebingungan proaktif yang diusulkan dapat dengan mudah diintegrasikan ke dalam protokol manajemen replika yang cocok untuk lingkungan yang bermusuhan. Ini dapat menggabungkan semua pengurutan ulang alamat, bantalan tumpukan, pengurutan ulang panggilan sistem, pengacakan set instruksi, pengacakan heap, dan pengacakan data sebagai pertahanan terhadap contoh permusuhan dalam sistem operasi komersial.

Kegagalan replika dikategorikan menjadi replika yang rusak dan disusupi. Replika yang disusupi berada di bawah kendali musuh. Ini disebut kegagalan Bizantium dalam literatur sistem operasi yang toleran terhadap kesalahan. Model kegagalan memiliki ambang batas kompromi yang membatasi jumlah replika yang dikompromikan. Namun, musuh yang

memiliki akses ke executable yang dikaburkan dapat melanggar ambang batas kompromi dengan menghasilkan serangan khusus yang pada akhirnya membahayakan semua replika. Untuk menghindari serangan permusuhan seperti itu, kebingungan proaktif akan me-reboot replika di seluruh periode yang konfigurasinya berubah seiring waktu sesuai dengan MTD. Kerahasiaan data diberlakukan dengan menyimpan data terenkripsi di server dengan kunci per server yang berbeda.

Kemudian teknik kriptografi dapat dikembangkan untuk melakukan komputasi pada data terenkripsi tersebut. Namun, kebingungan proaktif tidak dapat bertahan melawan semua serangan penolakan layanan (DoS) yang dapat memenuhi sumber daya seperti jaringan yang tidak berada di bawah kendali replika. Untuk menjamin bahwa penyegaran replika terjadi dalam jumlah waktu yang terbatas, beberapa komponen arsitektur operasi dibuat sinkron dengan memenuhi properti sinkronisitas yang kuat seperti jam kecepatan terbatas, prosesor sinkron, tautan tepat waktu, saluran reboot, dll. Mekanisme kebingungan proaktif dijelaskan dalam sistem terdistribusi nyata untuk memunculkan musuh terbatas yang mengevaluasi teknologi kebingungan.

Algoritma Pembelajaran Mendalam Manipulasi Game Theoretical dalam Aplikasi Keamanan Siber

Sekarang, kami mengalihkan perhatian pada aplikasi keamanan siber dari pembelajaran mendalam adversarial teoretis permainan. Liang dkk. meninjau kerangka teoritis permainan untuk menangani serangan jaringan dengan sistem deteksi intrusi (IDS) dan sistem pencegahan intrusi (IPS). Skenario aplikasi dikategorikan sebagai analisis serangan-pertahanan dan pengukuran keamanan. Model permainannya dirangkum menjadi permainan kooperatif dan permainan non kooperatif. Di sini, IDS menganalisis serangan siber dengan metode seperti identifikasi tanda serangan, deteksi pola, dan analisis statistik. Administrator jaringan bertindak sebagai pembela dalam pembelajaran mendalam yang bermusuhan.

Laporan pengukuran keamanan dibuat atas pengukuran keamanan jaringan seperti evaluasi kerahasiaan, integritas, ketersediaan, kerentanan, dan risiko keamanan dalam jaringan. Teori permainan digunakan dalam analisis serangan-pertahanan dari penilaian risiko dan pengambilan keputusan berdasarkan data dalam aplikasi jaringan dengan model permainan sinyal untuk memprediksi tindakan penyerang dan menentukan keputusan para pembela HAM. Berdasarkan jumlah tahapannya, model permainan dibedakan menjadi permainan statis/strategis yaitu permainan satu pukulan dengan informasi yang tidak sempurna, permainan dinamis/ekstensif dengan beberapa tahapan dan gerakan yang berhingga atau tidak terbatas, dan permainan stokastik menurut transisi.

Matriks probabilitas antara negara bagian tempat pemain mengambil tindakan dan menerima hadiah. Elemen dasar keseimbangan teoritis permainan yang akan didefinisikan oleh perancang algoritma adversarial adalah pemain, tindakan, hasil, dan strategi. Manshaei dkk. mengatur keamanan jaringan komputer ke dalam keamanan lapisan fisik dan MAC, jaringan yang mengatur dirinya sendiri, sistem deteksi intrusi, anonimitas dan privasi, ekonomi keamanan jaringan, dan kriptografi. Analisis keseimbangan teoritis permainan dan

desain mekanisme keamanan disajikan untuk setiap kategori keamanan dari masalah yang muncul dalam jaringan komputer. Agen jaringan dihitung sebagai individu, perangkat atau perangkat lunak, dan pengambil keputusan yang dapat bertindak secara kooperatif, egois, atau jahat.

Keputusan keamanan berdasarkan pendekatan teori permainan menghasilkan solusi keamanan dan privasi untuk alokasi sumber daya yang terbatas, keseimbangan risiko yang dirasakan, dan desain mekanisme insentif yang mendasari. Bergantung pada informasi tentang jenis musuh yang tersedia bagi pengambil keputusan, permainan keamanan mendukung pengambilan keputusan formal, pengembangan algoritma, dan pembelajaran mesin permusuhan dalam memprediksi perilaku penyerang dan interaksi antara penyerang dan pembela dalam hal ruang tindakan dan tujuan keputusan. -pembuat. Formulasi permainan keamanan dapat bervariasi antara permainan deterministik sederhana, permainan stokastik kompleks, dan permainan informasi terbatas dalam area fokus aplikasi seperti jamming dan penyadapan dalam jaringan nirkabel, sistem deteksi intrusi kolaboratif, privasi lokasi kooperatif, pemilihan jalur tor, kriptografi dalam multi-pihak.

Komputasi, pencabutan jaringan ad hoc seluler, dan keamanan jaringan kendaraan. Permainan keamanan kemudian dirumuskan dalam bentuk permainan dilema tahanan berulang yang terbatas, permainan pemimpin-pengikut Stackelberg, permainan fuzzy, permainan zero-sum berulang, permainan jumlah umum stokastik, permainan dinamis stokastik bukan jumlah nol, permainan stokastik Stackelberg dua pemain. permainan, permainan Bayesian, permainan koalisi, permainan biaya variabel, permainan informasi yang tidak lengkap, permainan fiktif, permainan pembicaraan murahan, dll. Analisis keseimbangan dari permainan tersebut kemudian memberikan wawasan analitik terhadap keputusan mengenai isu-isu seperti investasi keamanan dan manajemen patch dalam sistem jaringan yang kompleks.

Roy dkk. menyajikan taksonomi dunia maya untuk mengklasifikasikan jenis permainan yang digunakan dalam mekanisme pertahanan keamanan jaringan generasi mendatang dan komputasi aman. Permainan statis dianalisis berkenaan dengan informasi tidak sempurna yang lengkap dan informasi tidak sempurna yang tidak lengkap. Permainan dinamis dianalisis berkenaan dengan informasi sempurna yang lengkap, informasi sempurna yang lengkap, informasi sempurna yang tidak lengkap, dan informasi tidak sempurna yang tidak lengkap.

Otrok dkk. mengusulkan model pembelajaran teoretis permainan untuk sistem deteksi intrusi berbasis host (HIDS) untuk mengimbangi biaya komputasi yang tinggi dalam pembuatan dan deteksi alarm palsu untuk sistem dengan sumber daya terbatas seperti perangkat seluler nirkabel. Permainan informasi yang dinamis, non-kooperatif, multi-tahap, dan tidak lengkap diformulasikan untuk jaringan ad hoc seluler (MANET) menurut model probabilistik Bayesian dan model probabilistik Dempster-Shafer untuk representasi matematis dari ketidakpastian dan manajemen risiko di mana identitas penyerangnya tidak diketahui.

Konsep solusi permainan menentukan nilai fungsi keyakinan posterior pengguna untuk menentukan perilaku buruk dengan mengurangi kesalahan positif, meningkatkan

akurasi deteksi penyerang, dan mengoptimalkan efisiensi konsumsi sumber daya di HIDS. Ekuilibrium Bayesian sempurna menghitung serangkaian strategi yang optimal sehubungan dengan estimasi keyakinan yang dianggap sebagai probabilitas. Setelah pengukuran keyakinan mencapai ambang risiko yang telah ditentukan, HIDS dapat memutuskan apakah pengguna adalah penyerang atau bukan. Pengukuran keyakinan diperoleh dari bukti-bukti yang diamati dari suatu sumber data. Algoritme fusi keyakinan menggabungkan pengukuran keyakinan untuk menghasilkan keyakinan akhir dari domain masalah.

Dengan mengorbankan sumber daya komputasi tambahan, pengukuran keyakinan akhir seperti itu lebih tepat dibandingkan kemungkinan posterior Bayesian. Elemen permainan HIDS adalah ruang pemain dan tipe, ruang strategi, keyakinan sebelumnya, fungsi utilitas, dan tingkat deteksi HIDS. Nguyen dkk. membahas permainan keamanan Stackelberg zero-sum yang mengoptimalkan metode oracle ganda pada ruang tindakan yang besar secara eksponensial untuk mengalokasikan sumber daya deteksi botnet dalam solusi teoretis permainan untuk kebijakan pertahanan. Dua skenario eksfiltrasi data botnet diusulkan untuk mewakili vektor serangan jalur tunggal dan ganda untuk mencuri data jaringan sensitif.

Program linier bilangan bulat campuran mengoptimalkan respons terbaik dari pembela dan penyerang. Heuristik serakah memperkirakan dan mengimplementasikan ramalan. L'Huillier dkk. memanfaatkan permainan dinamis informasi yang tidak lengkap dalam deteksi penipuan phishing seperti penipuan email untuk mendapatkan informasi pribadi. Mesin vektor dukungan margin tertimbang bertindak sebagai pengklasifikasi permusuhan untuk pemfilteran phishing berbasis konten. Pemfilteran phishing pada aliran data pesan didasarkan pada algoritma online, algoritma pembelajaran generatif, dan algoritma pembelajaran diskriminatif berdasarkan pembelajaran mendalam adversarial teori permainan.

Di sini, phishing dapat dikategorikan menjadi phishing yang menipu dan phishing malware. Phishing yang menipu pada gilirannya dikategorikan menjadi rekayasa sosial, mimikri, spoofing email, penyembunyian URL, konten tak terlihat, dan konten gambar. Ekuilibrium Bayesian sempurna diusulkan sebagai konsep solusi permainan sinyal klasifikasi adversarial. Para pemain berperilaku sesuai dengan gagasan rasionalitas sekuensial di mana kumpulan informasi dalam permainan bentuk ekstensif menentukan keyakinan Bayesian tentang strategi keseimbangan yang didefinisikan dalam istilah strategi optimal gabungan untuk setiap agen serta keyakinan untuk setiap agen pada setiap kumpulan informasi di mana agen harus bergerak. Kemudian persyaratan keamanan sinyal diterapkan pada desain pengklasifikasi yang memperhitungkan strategi pertahanan optimal yang ditentukan oleh distribusi probabilitas atas tindakan pengklasifikasi dalam game.

Tindakan musuh kemudian memaksimalkan fungsi utilitas sesuai dengan kebutuhan sinyal. Keyakinan musuh mengikuti aturan Bayesian. Pemecah program kuadrat mengimplementasikan algoritma online untuk optimasi minimal berurutan dalam mesin vektor dukungan margin tertimbang. Korpus phishing dianalisis untuk mengetahui sifat struktural bagian isi pesan, analisis tautan di sekitar alamat IP dalam pesan, elemen pemrograman seperti HTML, formulir JavaScript yang digunakan dalam pesan, dan ambang

batas spam yang direkomendasikan. Frekuensi daftar kata dan fitur pengelompokan yang mewakili strategi phishing digunakan sebagai masukan ke pengklasifikasi adversarial.

Nagurney dkk. mengembangkan model teori permainan jaringan rantai pasokan antara pengecer dan pasar permintaan untuk memaksimalkan keuntungan yang diharapkan. Fungsi biaya investasi keamanan siber dan transaksi produk yang optimal direkomendasikan sehubungan dengan batasan anggaran non-linier. Preferensi konsumen ditentukan oleh fungsi harga permintaan yang menunjukkan permintaan produk dan tingkat keamanan siber di jaringan rantai pasokan. Kerentanan jaringan rantai pasokan serta pengecer pesaing dirumuskan sebagai masalah ketimpangan yang bervariasi. Ekuilibrium Nash menemukan ekspresi optimal untuk transaksi produk, tingkat keamanan, dan batasan anggaran. Analisis sensitivitas dengan contoh-contoh permusuhan mengukur perubahan anggaran, perubahan fungsi harga permintaan, kerugian finansial pada transaksi produk, dan biaya reputasi pada investasi keamanan siber.

Wang dkk. menyajikan model spoofing berbasis agen untuk manipulasi harga di pasar keuangan. Di sini, pedagang manusia bekerja melalui buku pesanan terbatas (limit order book) yang berisi informasi pribadi dan observasi rumit tentang lingkungan pasar yang kompleks untuk instrumen keuangan. Agen teoretis permainan kemudian mengikuti dua strategi perdagangan yang berbeda. Strategi zero intelijen (ZO) yang tidak dapat dipalsukan mengabaikan buku pesanan. Ini bertindak sebagai dasar pemodelan. Pembelajaran keyakinan heuristik yang dapat dimanipulasi (HBL) mengeksploitasi buku pesanan untuk memprediksi hasil harga. HBL diterapkan pada lingkungan pasar yang kompleks dalam siklus penuh pesanan dengan pesanan yang persisten, gabungan nilai-nilai privat dan fundamental, observasi yang berisik, kedatangan stokastik, dan kemampuan untuk memperdagangkan banyak unit dengan fleksibilitas beli atau jual.

Analisis teoretis permainan pada imbalan agen yang disimulasikan dihitung di lingkungan yang berbeda secara parametrik untuk model pasar guna mengukur pengaruh spoofing terhadap kinerja pasar menurut konsep solusi yang ditemukan dalam keseimbangan strategis. Strategi spoofing sederhana mengenai perilaku perdagangan dan efisiensi pasar terbukti menyesatkan pedagang, mendistorsi harga, dan mengurangi total surplus. Pedagang HBL juga bisa mendapatkan keuntungan dari penemuan harga dan kesejahteraan sosial dalam pembelajaran mesin teoritis permainan. Rancangan mekanisme pasar diusulkan untuk mendisinsentifkan manipulasi. Kemudian variasi strategi perdagangan diusulkan untuk meningkatkan ketahanan pembelajaran dari informasi pasar. Oleh karena itu, pembelajaran mendalam yang bersifat adversarial berpotensi mengubah lanskap pasar keuangan (seperti pasar valuta asing dan komoditas) dari ekosistem keputusan manusia menjadi teknologi perdagangan algoritmik dengan model pasar berbasis agen teoretis permainan.

Platform perdagangan otomatis tidak hanya dapat meningkatkan efisiensi pasar tetapi juga meningkatkan risiko pasar dan fluktuasi pasar karena praktik manipulatif seputar kerentanan yang didorong oleh algoritma. Spoofing didefinisikan sebagai penyerahan sejumlah besar pesanan beli/jual palsu dengan maksud untuk membatalkannya sebelum dieksekusi, sehingga merusak sinyal limit order book mengenai penawaran dan permintaan.

Pesanan spoof biasanya ditempatkan di luar harga terbaik saat ini untuk menyesatkan investor sebelum pergerakan pasar dapat memicu perdagangan. Evaluasi eksperimental berkisar pada model pasar lelang ganda yang berkelanjutan dengan satu sekuritas yang diperdagangkan.

Mekanisme pasar dirancang untuk memiliki elemen kunci dari kondisi struktur mikro pasar seperti guncangan fundamental dan gangguan observasi. Spoofing buku pesanan batas ditafsirkan sebagai serangan waktu pengambilan keputusan pada model pembelajaran mesin untuk menghasilkan contoh permusuhan dengan batasan domain pada aliran pesanan. Perilaku keseimbangan pasar kemudian ditentukan oleh aspek pembelajaran mendalam adversarial teori permainan di sekitar konteks pasar untuk menyeimbangkan kekuatan dan kemanjuran pembelajaran mesin dari informasi pesanan dengan mekanisme penyelubungan. Pekerjaan terkait adalah regresi linier permusuhan dengan banyak pelajar oleh Tong et al..

Nisioti dkk. menyajikan kerangka pendukung keputusan berbasis data yang disebut DISCLOSE untuk mengoptimalkan investigasi forensik terhadap pelanggaran keamanan siber. DIS-CLOSE mengelola gudang informasi intelijen ancaman yang berisi spesifikasi taktik, teknik, dan prosedur (TTP). TTP permusuhan diperoleh untuk serangan kompleks dengan beberapa jalur serangan mulai dari wawancara profesional keamanan siber, repositori MITRE ATT&CK STIX, dan Common Vulnerability Scoring System (CVSS). Di sini, pembelajaran mendalam permusuhan teoretis permainan bertindak sebagai hipotesis penalaran yang meningkatkan efisiensi penyelidikan forensik dengan mengurangi waktu dan sumber daya untuk proses penalaran yang kuat tentang hubungan logis antara bukti yang ditemukan secara objektif.

Penalaran strategis teoretis permainan dapat dikontraskan dengan kerangka penalaran dalam pembelajaran mesin seperti penalaran berbasis kasus, penalaran berbasis aturan, penalaran berbasis data, dll. Di sini, penalaran berbasis aturan dibingkai dari kombinasi aturan, model, dan model yang telah ditentukan sebelumnya. dan data sebelumnya. Hubungan probabilistik antara tindakan penyerangan yang tersedia, temuan investigasi forensik, manfaat dan biaya setiap inspeksi, dan anggaran yang tersedia bagi penyelidik dipertimbangkan dalam kerangka pendukung keputusan DISCLOSE. Liu dkk. melakukan survei sistematis tentang ancaman keamanan dalam pembelajaran mesin dari aspek teori pembelajaran pada fase pelatihan/penalaran dan pengujian/penyimpulan.

Penekanan khusus adalah pada penyimpangan distribusi data yang disebabkan oleh sampel yang bermusuhan dan pelanggaran informasi sensitif berikutnya dalam algoritma pembelajaran mesin statistik. Kemampuan permusuhan untuk menciptakan manipulasi permusuhan sesuai dengan tujuan permusuhan dikualifikasikan berdasarkan dampak ancaman keamanan penyebab atau eksplorasi, persentase data pelatihan dan pengujian yang dikendalikan oleh musuh, dan luasnya fitur dan parameter yang diketahui oleh musuh. Jenis serangan kemudian dikategorikan menjadi serangan kausatif, serangan eksplorasi, serangan integritas, serangan ketersediaan, serangan pelanggaran privasi, serangan bertarget, dan serangan sembarangan.

Teknik pertahanan pembelajaran mesin saat ini dikategorikan ke dalam mekanisme penilaian keamanan, tindakan pencegahan dalam fase pelatihan, tindakan pencegahan dalam

fase pengujian atau kesimpulan, keamanan data, dan privasi data. Mereka sangat penting dalam perancangan sistem cerdas yang belajar dari data besar dengan efisiensi tinggi, biaya komputasi minimum, dan akurasi prediksi atau klasifikasi yang wajar. Xue dkk. melakukan survei masalah keamanan dalam sistem pembelajaran mesin untuk merangkum pertahanan penanggulangan, teknik pembelajaran yang aman, dan metode evaluasi keamanan. Ancaman keamanan pembelajaran mesin dan model serangan dikategorikan ke dalam peracunan set pelatihan, pintu belakang di set pelatihan, serangan contoh permusuhan, pencurian model, dan pemulihan data pelatihan sensitif. Contoh permusuhan didefinisikan sebagai properti intrinsik model pembelajaran mendalam.

Overfitting model kemudian ditemukan memiliki pengaruh penting dalam pemulihan data pelatihan sensitif oleh musuh yang dapat melakukan serangan inferensi keanggotaan dan serangan inversi model. Model ancaman, pendekatan serangan, dan teknik pertahanan untuk sistem pembelajaran mesin dianalisis secara sistematis untuk menghasilkan intelijen ancaman siber di berbagai tahap rantai pembunuhan siber. Contoh serangan adversarial ditemukan pada pemfilteran spam email, deteksi malware Android, sistem otentikasi biometrik, sistem pengenalan wajah, pengenalan rambu jalan, pengenalan kamera ponsel, sistem kontrol suara, dan serangan objek 3D.

Mereka dikontraskan dengan serangan pintu belakang dan serangan Trojan untuk membuat data berbahaya untuk model target. Arahan penelitian di masa depan diberikan sebagai serangan dalam kondisi fisik nyata, teknik pembelajaran mesin yang menjaga privasi, perlindungan kekayaan intelektual (IP) berbasis watermarking pada jaringan saraf dalam, teknik keamanan pembelajaran mesin jarak jauh atau ringan, dan metode evaluasi keamanan pembelajaran mesin yang sistematis untuk menghasilkan alasan yang mendasari serangan dan pertahanan pada pembelajaran mesin. Serangan pembelajaran mendalam dapat dilakukan dengan model generatif mendalam seperti GAN untuk mematahkan kerangka pembelajaran terdistribusi atau gabungan. Pembelajaran mendalam adversarial teoretis permainan dikategorikan sebagai teknik pertahanan untuk mensimulasikan serangan, membuat strategi ketahanan, dan mendeteksi fitur abnormal dalam desain pengklasifikasi.

Hal ini kontras dengan teknik pertahanan lainnya seperti sanitasi data, deteksi anomali masukan, pemangkasan masukan dan penyempurnaan model, pelatihan ulang permusuhan, distilasi defensif, penyembunyian gradien, dan pengacakan masukan. Ini juga dapat dikombinasikan dengan teknik defensif untuk melindungi data sensitif seperti kriptografi, steganografi, kerangka pembelajaran mesin terdistribusi, platform tepercaya, dan pr/osesor. Evaluasi keamanan algoritme pembelajaran mesin juga dapat memperoleh manfaat dari kumpulan data pelatihan, pengujian, dan validasi yang dihasilkan oleh pembelajaran mendalam adversarial teoretis permainan dalam paradigma desain-untuk-keamanan, bukan paradigma desain-untuk-kinerja untuk pembelajaran mesin.

Kurva evaluasi keamanan juga dapat dibuat di sekitar ukuran kinerja untuk pembelajaran mesin dan fungsi biaya untuk pembelajaran mendalam permusuhan untuk mengkarakterisasi metrik evaluasi kinerja, ketahanan, keamanan, dan privasi sistem

pembelajaran yang dihitung dengan adanya berbagai jenis musuh yang memiliki kekuatan serangan dan pengetahuan yang berbeda. tingkat.

4.5 TEORI PERMAINAN YANG TANGGUH DALAM MANIPULASI PEMBELAJARAN MESIN

Dalam permainan Stackelberg, strategi manipulasi dimodelkan dan diselesaikan untuk alasan solusi dan masalah pengambilan keputusan yang mendefinisikan keseimbangan Nash. Ruang solusi untuk keseimbangan Nash dinyatakan dalam kondisi perlu dan cukup untuk kriteria konvergensi pemain game. Kriteria konvergensi yang umum adalah

i. permainan zero-sum vs non-zero sum game;
ii. permainan dua pemain vs multipemain;
iii. permainan statis vs permainan evolusi;
iv. permainan berurutan vs permainan berkelanjutan; dan
v. permainan deterministik vs permainan stokastik. Strategi umum para pemain mempertimbangkan kasus-kasus di mana sepasang pemain (i) tidak mengetahui kriteria kinerja masing-masing; (ii) menghitung strategi satu sama lain dengan kecepatan berbeda; (iii) memiliki fungsi pembayaran linier dan nonlinier yang mungkin terputus-putus atau tidak; dan (iv) berpartisipasi dalam permainan dengan kendali terdistribusi vs kendali terdesentralisasi. Dalam permainan seperti itu, strategi Stackelberg dan ekuilibria Nash dianalisis dalam kaitannya dengan sifat struktural matriks koefisien persamaan diferensial matriks-Riccati tingkat tinggi.

Optimalisasi fungsi pembayaran teoritis permainan tersebut menghadirkan masalah kompleks dalam teori optimasi. Masalah seperti ini sering dimodelkan sebagai masalah keputusan dalam permainan diferensial non-kooperatif. Solusi untuk masalah ini disajikan sebagai Pareto optima, keseimbangan Nash dan Stackelberg, dan solusi co-co (kooperatif-kompetitif) untuk fungsi pembayaran.

Persamaan diferensial Riccati juga dianalisis sebagai permainan diferensial dalam teori kendali optimal. Jika pemain teori permainan dapat mengamati keadaan sistem kendali, maka keseimbangan Nash dihitung berdasarkan solusi loop terbuka untuk sistem kendali. Jika pemain teori permainan tidak dapat mempertimbangkan strategi umpan balik, maka ekuilibrium Nash dihitung berdasarkan solusi loop tertutup untuk sistem kendali. Prinsip pemrograman dinamis digunakan sebagai metode komputasi untuk menemukan permainan optimal teoritis pada kondisi perlu dan cukup untuk sistem kendali optimal.

Selain itu, persamaan keadaan diferensial parsial dari sistem kontrol dapat menambah fungsi pembayaran pemain untuk menghasilkan kontrol stokastik dalam interaksi teoritis permainan. Di sini, keseimbangan teoretis permainan ditentukan oleh kondisi perlu dan cukup pada penyelesaian koefisien untuk persamaan diferensial, selisih, dan aljabar Stackelberg Riccati. Studi tentang keseimbangan tersebut dan metode komputasi numeriknya adalah subjek teori permainan evolusioner dan diferensial.

Zhou dkk. memodelkan beberapa jenis musuh dalam kerangka permainan Stackelberg bersarang. Seorang pembelajar dengan satu pemimpin harus menghadapi musuh dengan banyak pengikut. Solusi dari permainan tersebut adalah strategi campuran yang optimal bagi

pemimpin untuk bermain dalam permainan tersebut. Solusi solusi keseimbangan Stackelberg dua pemain digunakan sebagai strategi dalam permainan multipemain Bayesian Stackelberg. Permainan Stackelberg diselesaikan sebagai masalah pemrograman kuadratik bilangan bulat campuran (MIQP).

Ratliff dkk. mencirikan keseimbangan Nash dalam permainan berkelanjutan di ruang strategi non-cembung. Kondisi yang cukup diberikan untuk keseimbangan Nash diferensial. Mereka memerlukan evaluasi biaya pemain dan turunannya. Sudut pandang sistem dinamis diambil untuk menganalisis konvergensi strategi respons terbaik menuju keseimbangan yang stabil. Hasil dalam pemrograman non-linier dan kontrol optimal memberikan kondisi perlu dan cukup orde pertama dan kedua untuk optima lokal yang dinilai sebagai titik kritis dari fungsi bernilai nyata pada manifold data pelatihan. Permainan berkelanjutan seperti ini muncul dalam membangun manajemen energi, penetapan harga keamanan jaringan, optimalisasi waktu perjalanan dalam jaringan transportasi, dan integrasi energi terbarukan ke dalam sistem energi.

Model osilator berpasangan dipilih untuk mengilustrasikan properti sistem dari permainan berkelanjutan. Mereka memiliki aplikasi dalam jaringan listrik, jaringan lalu lintas, robotika, jaringan biologis, dan kontrol gerak terkoordinasi. Dianetti dkk. menyelidiki keberadaan kesetimbangan Nash dalam permainan diferensial stokastik pengikut monoton di mana setiap pemain memiliki biaya submodular. Masalah pengikut monoton melacak proses kontrol stokastik untuk mengoptimalkan kriteria kinerja. Ini memiliki aplikasi di bidang ekonomi dan keuangan, riset operasi, teori antrian, biologi matematika, teknik dirgantara, dan matematika asuransi.

Hal ini memungkinkan strategi umpan balik eksplisit untuk menghitung keseimbangan dalam strategi loop terbuka dan loop tertutup dalam permainan investasi yang tidak dapat diubah. Schuurmans dkk. memperkenalkan permainan pembelajaran mendalam. Optimalisasi model pembelajaran mendalam yang diawasi dinyatakan sebagai keseimbangan Nash dalam sebuah permainan. Sebuah bijeksi terbentuk antara keseimbangan Nash dari permainan gerak simultan dan titik KKT dari jaringan saraf asiklik terarah dalam pembelajaran mendalam. Kemudian algoritma pencocokan penyesalan bebas langkah diusulkan untuk pelatihan stokastik untuk menghasilkan model pembelajaran terawasi jarang dalam pembelajaran mendalam.

Dengan demikian, pembelajaran yang diawasi direduksi menjadi bermain game. Permainan gerak simultan satu pukulan didefinisikan untuk masalah pembelajaran satu lapis. Minimisasi penyesalan juga dapat menguraikan game multipemain menjadi beberapa game dua pemain. Lippi menggunakan kerangka pembelajaran relasional statistik (SRL) dalam deskripsi dan analisis permainan. SRL menggabungkan logika orde pertama dengan model grafis probabilistik untuk menangani ketidakpastian data dan ketergantungan representasinya. SRL dapat digunakan dalam permainan seperti permainan informasi parsial, permainan grafis, dan permainan stokastik.

Algoritme inferensi dalam SRL seperti propagasi keyakinan atau rantai Markov Monte Carlo dapat digunakan untuk pemodelan lawan, menemukan keseimbangan Nash, dan

menemukan solusi optimal Pareto. SRL menghasilkan klausa logika probabilistik untuk mendeskripsikan strategi dalam permainan sebagai formalisme tingkat tinggi yang dapat ditafsirkan manusia. Permainan digambarkan sebagai domain kepentingan, strategi, aliansi, aturan, hubungan, dan ketergantungan antar pemain. Teknik dari pemrograman logika induktif kemudian dapat mengekstraksi aturan dari basis pengetahuan predikat logika yang membantu penalaran probabilistik dalam pengambilan keputusan berdasarkan data. SRL juga dapat dikombinasikan dengan teori permainan untuk mempelajari struktur model dari data.

Beberapa metodologi SRL yang cocok untuk penalaran strategis dalam teori permainan adalah Causal Probabilistic Time Logic (CPT-L), Logical Markov Decision Programs (LOMDPs), DTProbLog, Infinite Hidden Relational Trust Model (IHRTM), Infinite Relational Model menuju pembelajaran kepercayaan, Relasional Pembelajaran Penguatan, Logika Lunak Probabilistik, dan Logika Pilihan Independen (ICL). SRL juga dapat menangani ketidakpastian data untuk menangani model game dengan informasi yang tidak lengkap atau tidak diketahui. Dalam pembelajaran mesin, SRL memiliki aplikasi untuk klasifikasi kolektif seperti prediksi tautan, klasifikasi objek, dan deteksi grup dalam domain aplikasi seperti jejaring sosial, bioinformatika, kemoinformatika, pemrosesan bahasa alami, dan web semantik. Inferensi maksimum a posteriori (MAP) difasilitasi oleh SRL untuk pembelajaran mendalam adversarial teori permainan dengan ekuilibria Nash dan Pareto optima dalam permainan strategis yang mengandung logika Markov.

Bektor dkk. menulis teks tentang permainan fuzzy. Ini mencakup teori keputusan fuzzy dan pemrograman matematika fuzzy yang memiliki penerapan dalam pembelajaran mendalam teori permainan yang beroperasi dalam lingkungan pembelajaran fuzzy. Bonanno dkk. membahas epistemik teori permainan. Hal ini berguna dalam mendefinisikan rasionalitas dan penalaran tipe musuh dalam pembelajaran mendalam permusuhan teoritis permainan. Keyakinan probabilistik dan preferensi utama para pemain dapat dianalisis baik secara semantik maupun sintaksis dalam kaitannya dengan pengakuan timbal balik atas rasionalitas di antara mereka.

Tsipras dkk. membuktikan bahwa ada trade-off antara ketahanan yang berlawanan dan kinerja pembelajaran dalam desain pengklasifikasi yang kuat. Biaya yang diakibatkannya menyebabkan metode pelatihan yang mahal secara komputasi dalam pembelajaran adversarial. Ketahanan permusuhan didefinisikan dalam kaitannya dengan rendahnya nilai perkiraan kerugian permusuhan. Pelatihan yang sangat kuat bertindak sebagai augmentasi data untuk mengatur model pembelajaran dan menghasilkan solusi analitik yang lebih baik. Terdapat hubungan erat antara ketahanan adversarial dan kompleksitas sampel dari pembelajaran kuat yang menggunakan asumsi generatif pada data.

Model pembelajaran yang kuat memiliki interpolasi fitur yang bersih serupa dengan yang diperoleh dari pembelajaran generatif mendalam. Kamu dkk. membahas kompresi model yang menjaga ketahanan permusuhan dengan pelatihan permusuhan dan pemangkasan beban secara bersamaan. Hal ini dapat digunakan untuk skenario pembelajaran mendalam yang kritis terhadap keamanan dalam sistem tertanam dengan sumber daya terbatas seperti ponsel, perangkat IoT, perangkat kesehatan pribadi, mengemudi otonom,

sistem udara tak berawak, dll. Pemangkasan bobot mengeksploitasi ketersebaran dalam jaringan saraf dalam untuk memangkas bobot koneksi tanpa penurunan kinerja yang nyata.

Model yang dilatih secara musuh terlihat lebih jarang dibandingkan model yang dilatih secara alami. Selain itu, pelatihan permusuhan memerlukan kapasitas jaringan yang lebih besar dibandingkan untuk mencapai kekuatan permusuhan yang kuat dibandingkan hanya mengklasifikasikan contoh-contoh yang tidak berbahaya dengan benar. Pemangkasan bobot berbasis ADMM (metode pengali arah bolak-balik) diusulkan. Merek dkk. memperkenalkan permainan iterasi pencocokan sen (IMP) untuk menganalisis kemampuan belajar permusuhan, kemampuan belajar konvensional, dan kemampuan perkiraan. “Kelengkapan” kemampuan belajar dibahas dalam kaitannya dengan identifikasi bahasa dan konsistensi statistik dalam batasnya. Metode prediksi apa pun yang dapat dihitung tidak mungkin lengkap. Mesin Turing dapat mempelajari bahasa komputasi apa pun. Gagasan tentang kemampuan belajar seperti itu penting dalam statistik, ekonometrik, pembelajaran mesin, inferensi induktif, dan penambangan data. Dalam konteks ini, pembelajaran mendalam adversarial teori permainan dapat digunakan dalam masalah kemampuan belajar dan perkiraan ketika pemain mencoba untuk belajar satu sama lain melalui observasi.

Eksistensi dan Keunikan Solusi Ekuilibrium Teoritis Permainan

Nash dkk. adalah makalah asli yang mendefinisikan konsep solusi keseimbangan dalam game multipemain dengan strategi murni. Strategi campuran kemudian menjadi distribusi probabilitas atas strategi murni. Nash dkk. membahas teori permainan non-kooperatif tanpa koalisi. Setiap pemain bertindak secara independen dan bukan sebagai koalisi tanpa komunikasi atau kolaborasi dengan pemain lain. Konsep solusi keseimbangan kemudian merupakan generalisasi dari konsep permainan zero-sum dua pemain.

Beberapa konsep solusi dikembangkan untuk memenuhi hipotesis pembelajaran dalam pembelajaran mendalam adversarial teoritis permainan seperti solusi bentuk geometris dan analisis kontradiksi pada strategi keseimbangan. Transferabilitas dan komparabilitas antara fungsi-fungsi hasil yang berlawanan juga merupakan sebuah pertanyaan untuk membandingkan solusi keseimbangan teoritis permainan dalam aplikasi dunia nyata. Sistem dinamis permainan non-kooperatif dapat dikembangkan untuk mereduksi permainan kooperatif dengan negosiasi pra-permainan dalam permainan kooperatif yang menjadi permainan dalam permainan non-kooperatif untuk menggambarkan semua imbalan pemain dalam permainan yang tidak terbatas.

Medanik dkk. mengembangkan ekspresi eksplisit dari strategi Stackelberg multilevel loop terbuka untuk kontrol dalam masalah pengambilan keputusan sekuensial deterministik. Sistem linier kontinu diselesaikan dengan kriteria optimasi kuadrat untuk mengkarakterisasi kontrol Stackelberg. Persamaan diferensial Riccati matriks persegi tingkat tinggi juga diformulasikan untuk mengkarakterisasi kontrol Stackelberg dengan matriks koefisien dalam sistem dinamis yang digunakan untuk inferensi statistik dari sifat strukturalnya. Strategi yang dipilih dalam urutan pengambilan keputusan untuk seorang pemain tersedia untuk pemain lain setelah permainan saat ini. Dimensi sistem dinamik terkait mewakili batasan diferensial untuk menentukan strategi optimal pengendalian berikutnya dalam urutan pengambilan

keputusan yang dihasilkan oleh subsistem dalam sistem pembelajaran yang saling berhubungan.

Freiling dkk. mempelajari keberadaan dan keunikan keseimbangan Stackelberg dalam permainan diferensial dua pemain dengan struktur informasi loop terbuka. Kondisi keberadaan yang memadai diturunkan untuk kesetimbangan loop terbuka untuk menyelesaikan persamaan diferensial matriks Riccati. Parameter invarian waktu dibahas untuk mengatasi penyimpangan konsep dalam solusi kesetimbangan. Solusi keseimbangan dapat diperluas dengan teori permainan non-kooperatif yang memiliki struktur hierarki, fungsi biaya, dan pola informasi data sampel yang berbeda dalam permainan diferensial Stackelberg. Persamaan diferensial linier menjelaskan batasan vektor keadaan dalam permainan. Fungsi pembayaran dibangun untuk memenuhi kondisi perlu dan cukup pada konsep solusi yang diperoleh dengan menyelesaikan persamaan diferensial.

OReilly dkk. mempelajari dinamika musuh siber untuk memperkuat pertahanan siber berdasarkan kriteria ketahanan siber. Dinamika tersebut dirumuskan sebagai sistem ko-evolusi kompetitif yang menghasilkan banyak perlombaan senjata untuk menghasilkan solusi yang kuat. Proses ko-evolusi ini dirancang dalam konteks skenario keamanan siber jaringan di mana pembela HAM memanfaatkan kecerdasan buatan (AI) untuk mendapatkan keunggulan kompetitif dalam lingkungan persaingan yang asimetris. AI Adversarial menerapkan konfigurasi trade-off defensif (kasus terburuk, kasus rata-rata) untuk mengantisipasi berbagai kemungkinan perilaku permusuhan dengan mengacu pada dampak, sasaran, strategi, atau taktik yang diharapkan. Dampak yang diharapkan dapat berupa kombinasi biaya finansial, tingkat gangguan, atau risiko hasil. Peringkat untuk konfigurasi defensif dihasilkan dari metode pencarian stokastik untuk mengeksplorasi ruang strategi dalam simulasi keterlibatan kompetitif dari perilaku musuh. Keterlibatan kompetitif terjadi antara populasi yang bermusuhan yang menjalani seleksi berdasarkan kinerja dan variasi untuk beradaptasi.

Logika ko-evolusioner kemudian menghasilkan dinamika permusuhan di seluruh populasi di mana pihak-pihak yang bermusuhan terlibat dan mengukur hasil mereka dengan mengacu pada musuh-musuh lainnya. Konfigurasi pertahanan yang kuat dibentuk oleh algoritma ko-evolusi yang membantu menghasilkan perilaku yang beragam. Keberagaman perilaku diukur dengan “konsep solusi” dari ketahanan permusuhan. Ada anggaran perhitungan atau waktu yang tetap untuk setiap keterlibatan yang bermusuhan. Kasus penggunaan simulasi-pemodelan pada ancaman permusuhan dan model pertahanan dalam keamanan komputer kemudian mendukung emulasi pembelajaran mendalam permusuhan dengan rincian model yang bervariasi. Tingkat granularitas tersebut mencakup serangan penolakan layanan dalam jaringan peer-to-peer, penyusupan perangkat dalam jaringan perusahaan, dan pertahanan yang menipu terhadap pengintaian internal musuh dalam jaringan yang ditentukan perangkat lunak.

Kriteria konvergensi algoritma pencarian co-evolusi juga memfasilitasi visualisasi dan perbandingan perilaku permusuhan dalam dinamika permusuhan generatif. Dalam eksperimen, strategi penyerang dan pembela digabungkan dalam konteks simulasi pertahanan jaringan komputer abstrak yang tunggal dan abstrak. Estimasi proses Gaussian

memperkirakan keterlibatan permusuhan yang tidak pasti. Teknik pemberi rekomendasi digunakan untuk memperkirakan fungsi kebugaran lawan. Grid spasial digunakan untuk mengurangi ruang pencarian pada keterlibatan kepentingan berpasangan. Keterlibatan permusuhan disimpan dalam cache agar berfungsi dalam anggaran evaluasi kesesuaian waktu yang tetap.

Lingkungan keterlibatan mendukung pengujian jaringan, simulator, dan model yang spesifik masalah. Urutan tindakan perilaku untuk menyerang dan bertahan diungkapkan dengan tata bahasa bentuk Backus-Naur (BNF) yang merupakan representasi tata bahasa bebas konteks dari perilaku permusuhan sebagai aturan pengambilan keputusan berdasarkan data. Tata bahasa BNF mengkomunikasikan fungsionalitas pembelajaran mendalam yang bermusuhan dan memungkinkan perjalanan percakapan dan validasi model dalam domain aplikasi wacana. Tata bahasa BNF, lingkungan keterlibatan, dan fungsi kebugaran bervariasi menurut tipe musuh. Modularitas dan kegunaannya kembali menghasilkan rekayasa perangkat lunak yang efisien dan keunggulan pemecahan masalah. Fungsi kesesuaian ko-evolusi kompetitif dan konsep solusi bergantung pada konteks keterlibatan yang bermusuhan. Ringkasan solusi "terbaik" diperoleh dari proses pemeringkatan dan penyaringan terhadap konsep-konsep solusi.

Cotter dkk. membahas teori optimasi terbatas yang diterapkan dalam klasifikasi Neyman-Pearson, optimasi yang kuat, dan pembelajaran mesin yang adil. Permainan non-zero sum dua pemain diselesaikan untuk memodelkan parameter optimasi dengan batasan yang tidak dapat dibedakan pada tingkat konvergensi dan jumlah kerugian/proporsi minimalisasi penyesalan. Masalah keadilan pembelajaran mesin dirumuskan sebagai minimalisasi kerugian empiris yang tunduk pada batasan keadilan yang bergantung pada data dan tidak dapat dibedakan. Jaringan saraf dalam dengan fungsi tujuan non-cembung digunakan untuk memodelkan masalah pembelajaran mesin. Batasan tersebut dinyatakan dengan fungsi variabel acak indikator. Kesetimbangan teoretis permainan kemudian menentukan pengklasifikasi stokastik.

Syrgkanis dkk. menyelidiki kesetimbangan yang berkorelasi dalam permainan bentuk normal multipemain yang tertanam dalam algoritma pembelajaran yang diatur dan pengurangan kotak hitamnya. Pembelajaran tanpa penyesalan digunakan untuk membuat keputusan pemain. Batasan penyesalan ditemukan di lingkungan permusuhan dengan algoritme tanpa penyesalan seperti bobot perkalian, penurunan cermin, dan mengikuti pemimpin yang diatur/terganggu. Dinamika tanpa penyesalan menghasilkan tingkat konvergensi yang lebih cepat untuk algoritma pembelajaran yang diatur. Pengurangan kotak hitam mereka dalam lingkungan teoritis permainan mempertahankan tingkat konvergensi sambil mempertahankan batas penyesalan atas ketahanan musuh. Hasilnya dibandingkan dengan permainan lelang simultan dalam hal utilitas, penyesalan, dan konvergensi menuju keseimbangan. Kesejahteraan dari permainan ini adalah hasil dari alokasi jumlah variabel dari imbalan dan pencocokan sumber daya yang terkait dengan masalah pencocokan bipartit yang tidak tertimbang.

Banyak permainan yang menarik dalam pembelajaran mendalam permusuhan berada di luar pemodelan dan penalaran yang dapat dilakukan. Wellman dkk. menyelidiki kesenjangan antara penalaran strategis dan teori permainan. Kompleksitas komputasi dalam penalaran otomatis terbukti disebabkan oleh jumlah agen, ukuran rangkaian strategi dan ruang kebijakan, tingkat informasi yang tidak lengkap dan tidak sempurna, dan perhitungan hasil yang diharapkan dalam lingkungan stokastik untuk pembelajaran permusuhan. Game empiris diperkenalkan sebagai simulator game yang melakukan penalaran strategis melalui simulasi interleaved dan analisis teoritis game. Landasan pembelajaran mendalam adversarial adalah skenario interaksi di mana informasi hasil diperoleh dari data dalam observasi dan simulasi. Membangun dan menalar tentang permainan empiris menyajikan sub-masalah menarik dalam simulasi, statistik, pencarian, dan analisis pembelajaran mendalam permusuhan teoritis permainan.

Formulasi permainan empiris didekomposisi menjadi parameterisasi ruang strategi atas rangkaian tindakan berkelanjutan atau multidimensi dan informasi tidak sempurna yang dikondisikan pada sejarah observasi. Parametrisasi ruang strategi dilakukan dengan strategi kandidat dalam struktur dasar atau kerangka untuk variasi parametrik pada arsitektur pencarian game seperti pengungkapan hasil yang sebenarnya, strategi respons terbaik yang rabun, dan pencarian pohon permainan dalam optimasi minimax dan max-max. Estimasi teknik permainan empiris yang tepat, ketat, dan otomatis direkomendasikan untuk dilakukan dengan teknik statistik seperti analisis Monte Carlo dalam pembelajaran aktif pilihan strategi berdasarkan jenis musuh, menyesuaikan faktor-faktor yang dapat diamati di dunia maya dengan efek yang diketahui pada hasil, menerapkan variasi kontrol pada mengukur imbalan yang disesuaikan dengan permintaan, pengurangan permainan hierarki untuk memengaruhi penghematan komputasi, kriteria teori informasi untuk memilih profil strategi, dan regresi dalam estimasi permainan untuk menggeneralisasi imbalan di ruang profil yang sangat besar berdasarkan data yang tersedia.

Vorobeychik dkk. menyelidiki permainan dengan strategi bernilai nyata di mana informasi hasil dipelajari pada sampel profil strategi. Masalah pembelajaran fungsi pembayaran dirumuskan sebagai masalah regresi standar dengan struktur yang diketahui dalam lingkungan multi-agen. Kinerja pembelajaran diukur sehubungan dengan kegunaan relatif dari strategi preskriptif daripada keakuratan fungsi hasil. Kegunaan relatif dari strategi juga diperkirakan sebagai target pembelajaran yang diawasi dan sebagai pemilih model pembelajaran. Dengan demikian, solusi keseimbangan teoretis permainan dapat ditentukan dari database pengalaman bermain game, bukan dari spesifikasi interaksi strategis dalam sebuah game. Menentukan solusi keseimbangan tersebut kemudian menjadi target pembelajaran mesin yang diterapkan pada game dengan rangkaian strategi yang sangat besar atau tak terbatas yang mendefinisikan agen berkelanjutan. Solusi keseimbangan dapat digunakan untuk merepresentasikan permainan multi-tahap sebagai permainan satu kali dengan rangkaian strategi yang merupakan fungsi dari semua kemungkinan sejarah permainan. Data hasil permainan diperoleh dari observasi agen lain yang memainkan permainan tersebut dan simulasi jalannya permainan secara hipotetis.

Tugas perkiraan fungsi pembayaran untuk penambangan data kemudian didefinisikan sebagai memilih fungsi dari kumpulan kandidat untuk meminimalkan ukuran penyimpangan dari fungsi pembayaran sebenarnya yang mewakili kotak hitam atau ramalan dalam pembelajaran mendalam yang bermusuhan. Regresi polinomial, regresi lokal, dan regresi mesin vektor pendukung digunakan untuk menghitung ekuilibria Nash murni. Fungsi yang dipelajari pada awalnya dibatasi pada subset strategi yang terbatas. Dinamika replikator mencari keseimbangan campuran simetris dengan algoritma evolusi berulang. Setelah sejumlah iterasi yang tetap, perkiraan keseimbangan Nash dalam matriks pembayaran pada subset strategi terpisah diperlakukan sebagai permainan yang dipelajari.

Oleh karena itu, metode pembelajaran regresi dan diskriminatif dalam pembelajaran yang diawasi menawarkan generalisasi pembelajaran mendalam permusuhan teoritis permainan ke ruang strategi yang tak terbatas di luar pengalaman yang tersedia secara langsung untuk pembelajaran mesin. Fungsi target yang mendukung penghitungan keseimbangan yang dapat dilakukan dapat dirumuskan dengan pembelajaran mendalam untuk mendukung “kemampuan belajar” fungsi hasil yang merugikan sesuai dengan pengorbanan pembelajaran yang ditemukan di berbagai lingkungan yang bermusuhan.

Penelitian kami mengenai trade-off pembelajaran mendalam permusuhan teoritis permainan antara kemampuan belajar dan kekuatan pembelajaran diskriminatif. Yang kami maksud dengan “kemampuan belajar” adalah kemampuan pengklasifikasi untuk memprediksi label yang benar (tanpa memperhatikan noise), dan yang dimaksud dengan ketahanan (robustness) adalah prediksi yang sama dengan atau tanpa noise (tanpa memperhatikan kebenarannya). Kerugian yang kami amati adalah bahwa semakin banyak kemampuan untuk dipelajari akan mengakibatkan semakin berkurangnya ketahanan dan sebaliknya. Tujuan utama dari penelitian kami adalah pengembangan algoritma pembelajaran mendalam adversarial teori permainan yang dapat diterapkan pada masalah penambangan data keamanan dunia maya.

Kami mengembangkan fungsi pembayaran teoretis permainan yang memodelkan batasan keputusan pembelajaran mesin yang diawasi. Mereka mengeksplorasi ketergantungan teori sistem antara pengacakan dalam manipulasi data permusuhan dan kemampuan generalisasi dalam optimasi pembelajar kotak hitam sehubungan dengan pembelajaran mendalam permusuhan teoritis permainan yang diusulkan. Pengoptimalan yang kuat seperti itu mempelajari teori ketahanan, keadilan, kemampuan menjelaskan, dan transparansi dalam pembelajaran mesin dengan permainan prediksi. Di sini, kami mengembangkan algoritma pembelajaran adversarial untuk keandalan, kemampuan belajar, efisiensi, dan kompleksitas dalam pembelajaran diskriminatif. Pembelajaran mendalam adversarial teoritis permainan yang dihasilkan diterapkan pada masalah klasifikasi dan optimasi dalam analisis data.

Hasil optimal dari kompleksitas sampel dalam formulasi teori permainan bergantung pada metode optimasi dan distribusi data target dari fungsi kerugian yang dihitung dalam kesetimbangan Nash dan Stackelberg. Kebijakan stokastik yang lebih baik dalam permainan keseimbangan teoretis akan menghasilkan solusi (kooperatif-kompetitif) dalam pemodelan

sistem (dinamik diferensial). Di sini, kita dapat membandingkan solusi pembelajaran adversarial teoretis permainan dengan dasar pembelajaran mesin seperti penyertaan noise dalam prosedur pengoptimalan, penyederhanaan lanskap fungsi dengan peningkatan ukuran model, skema untuk pengoptimalan stokastik bebas turunan, dan data. pengambilan sampel ulang dalam konteks algoritma pembelajaran variasional. Kami juga dapat mengeksplorasi strategi pengacakan untuk pengoptimalan yang kuat dalam game multipemain yang dapat diuraikan menjadi game prediksi atau game Stackelberg. Mereka sering dimodelkan sebagai masalah pengambilan keputusan dalam permainan diferensial non-kooperatif.

Teori Kontrol Optimal dan Teori Permainan yang Kuat

Huang dkk. menjelaskan permainan dinamis untuk desain sistem kontrol yang dapat diurai menjadi lapisan cyber, fisik, dan manusia. Masalah desain lintas lapisan menimbulkan tantangan keamanan dan ketahanan pada infrastruktur penting. Infrastruktur penting tersebut terlihat dalam sistem pengendalian industri di sektor-sektor seperti tenaga listrik, manufaktur, dan transportasi. Di sini, pandangan desain sistem kontrol mengambil perspektif penginderaan, kontrol, dan dinamika pembangkit yang terintegrasi dalam putaran umpan balik di lapisan fisik. Teknik desain kendali seperti kendali kuat, kendali adaptif, dan kendali stokastik menangani ketidakpastian informasi, gangguan fisik, dan gangguan yang merugikan dalam putaran umpan balik.

Kebisingan permusuhan terlihat di lapisan dunia maya dengan masalah komunikasi dan jaringan antara sensor dan aktuator serta di antara beberapa agen yang terdistribusi. Sebaliknya, lapisan manusia berkaitan dengan masalah pengawasan dan manajemen seperti koordinasi, operasi, perencanaan, dan investasi. Permasalahan pengelolaannya mencakup permasalahan sosial dan ekonomi, penetapan harga dan insentif, serta regulasi pasar dan analisis risiko. Dalam sistem otonom yang didukung cloud, kontrak layanan untuk layanan keamanan dapat mencakup kebijakan asuransi siber yang peka terhadap serangan dan kompatibel dengan insentif, yang dapat dirancang dengan pembelajaran adversarial teoritis permainan untuk memaksimalkan kesejahteraan sosial dan mengurangi bahaya moral.

Musuh mengeksploitasi permukaan serangan pada sistem kontrol untuk mengeksploitasi kerentanan zero-day dalam sistem otonom seperti kendaraan yang dapat mengemudi sendiri. Teori permainan memberikan kerangka kerja untuk interaksi strategis antar komponen dalam sistem yang kompleks untuk mengukur trade-off dari ketahanan, keamanan, dan ketahanan dalam kinerja sistem dalam lingkungan yang berlawanan dengan sistem kontrol. Dalam kerangka teori permainan, desain kontrol yang aman dan tangguh dipandang sebagai perpanjangan dari desain kontrol yang kuat. Area fokus aplikasi dihitung sebagai sistem otonom heterogen, permainan penipuan defensif untuk sistem kontrol industri, dan manajemen risiko jaringan cyber-fisik.

Tujuan dari sistem pengendalian yang tangguh adalah untuk mendapatkan jaminan kinerja dan mekanisme pemulihan ketika ketahanan dan keamanan gagal karena serangan musuh dan kegagalan sistem. Di sini, sistem kontrol yang kuat dapat menahan parameter dan gangguan yang tidak pasti karena paradigma pembelajaran mesin desain-untuk-keamanan. Mekanisme pertahanan dalam desain sistem kontrol yang kuat mencakup kriptografi, deteksi,

arsitektur solusi, dan protokol komunikasi. Desain sistem fisik diatur oleh persamaan aljabar diferensial untuk aplikasi seperti sistem robot multibody, sistem jaringan listrik, dan sistem distribusi air. Oleh karena itu, dinamika pembelajaran dapat digambarkan dengan proses keputusan Markov, persamaan perbedaan, dan persamaan diferensial parsial dalam pembelajaran mendalam adversarial teori permainan.

Literatur teori permainan dapat dikombinasikan dengan aspek kognitif, memori, komputasi, dan psikologis dari proses pengambilan keputusan manusia. Di sini, teori prospek menggabungkan keengganan terhadap kerugian dalam keputusan manusia dan membedakan persepsi kerugian dari kegunaan keuntungan. Mekanisme perhatian yang menggabungkan kognisi terbatas terhadap keputusan online manusia juga dapat dimasukkan ke dalam teori permainan. Kepemilikan layanan sistem kendali yang terdesentralisasi dapat menyediakan pembagian dan pemanfaatan sumber daya infrastruktur komputasi, komunikasi, dan penginderaan secara efektif. Kemudian implementasi dan investasi dalam pembelajaran mendalam yang bermusuhan untuk keamanan kemudian memungkinkan penawaran layanan berkualitas tinggi dengan memitigasi risiko keamanan pada tingkat layanan sistem kontrol.

Hal ini juga mencegah penyebaran risiko dunia maya secara real-time di berbagai skala sosio-ekonomi. Permainan dalam sistem otonom yang mempertimbangkan semua persyaratan keamanan tersebut dirancang untuk mencapai keseimbangan Gestalt Nash (GNE). GNE adalah konsep solusi keseimbangan teoritis permainan dimana tidak ada pemain yang mempunyai insentif untuk menyimpang dari solusi optimal. Solusi seperti itu dapat ditemukan tidak hanya dalam permainan modular yang ditentukan oleh interaksi agen-agen lokal, tetapi juga dalam permainan terintegrasi yang ditentukan oleh interaksi sistem-sistem global. Hal ini bertujuan untuk menemukan kemampuan beradaptasi diri, penyembuhan diri, dan ketahanan tangkas untuk sistem otonom heterogen yang menghadapi serangan tersembunyi multi-tahap seperti APT di lingkungan yang bermusuhan.

Prinsip pemrograman dinamis dapat digunakan untuk mempelajari sifat konvergensi permainan optima teoritis secara efisien. Dalam kerangka musuh variasional, pemodelan teoretis permainan juga dapat digunakan untuk memecahkan masalah pengambilan sampel dalam mekanisme privasi diferensial. Tugas pembelajaran mesin adalah menghasilkan sampel permusuhan dalam ruang laten metode variasional menurut distribusi yang ditentukan secara implisit yang penting untuk pengoptimalan dan klasifikasi manipulasi permusuhan dalam pembelajaran mesin. Kami kemudian dapat menyelesaikannya dengan menyajikan kerangka pembelajaran generatif yang mendalam dan skema pembelajaran diferensial dalam permainan privasi diferensial non-kooperatif untuk masalah pengambilan keputusan.

Lebih lanjut, permainan privasi diferensial dapat dianalisis dalam kaitannya dengan teori kontrol optimal. Jika pemain teori permainan dapat mengamati keadaan sistem kendali, maka keseimbangan Nash dihitung berdasarkan solusi loop terbuka untuk sistem kendali. Jika pemain teori permainan tidak dapat mempertimbangkan strategi umpan balik, maka ekuilibrium Nash dihitung berdasarkan solusi loop tertutup untuk sistem kendali. Prinsip pemrograman dinamis digunakan sebagai metode komputasi untuk menemukan permainan optimal teoritis pada kondisi perlu dan cukup untuk sistem kendali optimal. Selain itu,

persamaan keadaan diferensial parsial dari sistem kendali dapat menambah fungsi pembayaran pemain untuk menghasilkan kendali stokastik dalam interaksi teoritis permainan. Di sini, kesetimbangan teoretis permainan ditentukan oleh kondisi perlu dan cukup pada koefisien penyelesaian diferensial Stackelberg Riccati, selisih, dan persamaan aljabar kesetimbangan.

Studi tentang keseimbangan tersebut dan metode komputasi numeriknya adalah subjek teori permainan evolusioner dan diferensial. Kemudian kita akan menerjemahkan metode stokastik tersebut ke dalam bahasa pembelajaran permusuhan dengan musuh variasional. Peningkatan kapasitas pembelajaran, strategi pengacakan, dan fungsi pembayaran yang berpusat pada privasi dalam formulasi game akan memengaruhi regularisasi bobot dan batasan keputusan algoritme pembelajaran mesin yang disediakan sebagai layanan. Motif teori sistem dari pemrosesan sinyal non-linier dan statistik teori kontrol yang relevan untuk aplikasi penambangan data permusuhan juga dapat ditentukan dari pengetahuan domain. Dalam konteks ini, kami mengusulkan untuk mengeksplorasi dekomposisi wavelet dan pemodelan entropi maksimum dari distribusi data.

Grunwald dkk. mengembangkan teori keputusan dengan menghubungkan inferensi entropi maksimum dengan meminimalkan kemungkinan kerugian terburuk dalam permainan terbatas zero-sum antara pembuat keputusan pemain dan alam. Teori keputusan digunakan untuk memperoleh fungsi kerugian yang dapat digunakan dalam pembelajaran mendalam permusuhan. Distribusi entropi maksimum menentukan strategi minimax pengambil keputusan. Ukuran entropi relatif umum diperkenalkan untuk definisi teori keputusan tentang fungsi ketidaksesuaian dan kerugian. Entropi relatif umum sebanding dengan kerangka optimasi entropi lainnya seperti entropi Renyi dan informasi Fisher yang diharapkan. Franci dkk. menggambarkan pelatihan jaringan permusuhan generatif (GAN) sebagai masalah ketidaksetaraan variasional dengan solusi keseimbangan stokastik Nash.

Algoritme pelatihan maju-mundur santai stokastik diusulkan untuk GAN. Cai dkk. melakukan survei tentang kecanggihan GAN dari perspektif keamanan dan privasi. Strategi optimasi teoretis permainan di GAN digunakan untuk menghasilkan distribusi probabilitas multimodal berdimensi tinggi yang memiliki aplikasi penting dalam domain matematika dan teknik. Dalam metode pembelajaran mendalam permusuhan berbasis GAN, generator dapat digunakan tidak hanya untuk membuat contoh permusuhan tetapi juga merancang mekanisme pertahanan. Dalam penelitian privasi data, metode berbasis GAN dapat digunakan dalam steganografi gambar, anonimisasi gambar, dan pengkodean gambar. Jaringan permusuhan generatif variasi (VGAN) dan autoencoder variasional (VAE) dapat dibangun untuk mencapai keseimbangan antara privasi dan utilitas dalam gambar yang disintesis.

Dalam penelitian privasi model, metode berbasis GAN dapat digunakan untuk melindungi anonimisasi dan kebingungan privasi model pembelajaran. Dalam domain aplikasi pembelajaran mendalam permusuhan, GAN dapat menghasilkan contoh malware permusuhan dengan kompresi dan rekonstruksi data, pembuatan malware palsu, dan deteksi malware. Mereka dapat digunakan untuk membangun sistem bio-informasi untuk otentikasi;

masalah deteksi penipuan keuangan dalam penipuan kartu kredit, penipuan telekomunikasi, dan penipuan asuransi; deteksi botnet; dan deteksi intrusi jaringan.

Karakteristik kebisingan adversarial dapat didefinisikan sehubungan dengan gagasan kebisingan berikut dalam penelitian masa depan tentang pembelajaran mendalam adversarial teoretis permainan di mana model pembelajaran teoretis permainan melibatkan musuh evolusioner, musuh stokastik, dan musuh variasional yang menargetkan kinerja kesalahan klasifikasi jaringan saraf dalam dan saraf konvolusional. jaringan. Sejauh mana noise pada parameter model dan data pelatihan dapat bermanfaat bagi kualitas keseluruhan distribusi data yang dihasilkan oleh pembelajaran adversarial teoritis permainan bergantung pada proses noise adversarial tertentu dan sifat dari distribusi target yang dihasilkan.

- Kebisingan permusuhan berupa spam, outlier, diskontinuitas, dan merugikan
- Kebisingan yang bersifat permusuhan bukanlah kebenaran dasar, bukan sinyal, dan bukan merupakan sebuah kebohongan
- Kebisingan permusuhan adalah data palsu dan penemuan palsu
- Kebisingan permusuhan (adversarial noise) adalah prediksi dan kebocoran informasi yang tidak terduga
- Kebisingan permusuhan (adversarial noise) adalah kesalahan sisa dan objek yang tidak diketahui
- Kebisingan permusuhan merupakan kelas yang langka dan berstruktur jarang
- Kebisingan permusuhan adalah motif yang rumit dan keputusan yang salah
- Adversarial noise adalah contoh kesalahan klasifikasi dan nilai regresi yang salah
- Kebisingan permusuhan adalah sampel acak dan variabel laten
- Kebisingan permusuhan disebabkan oleh proses stokastik yang mendasarinya
- Kebisingan yang bersifat permusuhan tidak signifikan secara statistik
- Kegaduhan yang bersifat permusuhan tidak dapat dijelaskan

Ge dkk. merumuskan metode desain permainan untuk kontrol kuantum kuat yang dimainkan antara ketidakpastian (atau kebisingan) dan kontrol dalam perangkat keras kuantum. Lloyd dkk. memperkenalkan jaringan permusuhan generatif kuantum di mana generator dan diskriminator dilengkapi dengan pemroses informasi kuantum.

Romero dkk. memperkenalkan sirkuit kuantum variasional untuk meniru distribusi target. Sirkuit variasional untuk menyandikan informasi klasik ke dalam keadaan kuantum sangat berguna dalam aplikasi pembelajaran mesin seperti klasifikasi permusuhan. Kita dapat menjalankan algoritme pembelajaran komputasi yang mahal atau subrutin BLASnya secara efisien di komputer kuantum. Generalisasi kuantum dari pembelajaran mendalam adversarial pada distribusi data kuantum akan melibatkan pengambilan sampel kuantum, informasi kuantum, dan pemodelan kausalitas kuantum untuk menganalisis dekomposisi bias-varians dalam fungsi hasil adversarial yang dapat diterapkan ke dalam optimalisasi komputasi permainan prediksi acak.

Kita dapat memperoleh batasan utilitas untuk pembelajaran mendalam jaringan saraf kuantum dalam kerangka minimalisasi risiko empiris dan kerangka batasan kesalahan. Kita juga dapat mendefinisikan pembelajaran yang ditingkatkan kuantum melalui interaksi dalam

paradigma agen-lingkungan dari komputasi kuantum untuk memperoleh kriteria keterpisahan dalam mekanisme komputasi saraf dan kesalahan generalisasinya akibat pengukuran kuantum. Kita dapat mengembangkan teori kompleksitas sampel, verifikasi formal, dan automata fuzzy dalam model adversarial dengan jaminan andal yang diusulkan pada pembelajaran adversarial generatif kuantum dalam pelatihan dan pengoptimalan jaringan saraf kuantum. Kita dapat membuat pemroses informasi kuantum dengan tujuan khusus seperti annealer kuantum yang sangat cocok dengan arsitektur pembelajaran mendalam yang bermusuhan.

Skema pembelajaran klasik-kuantum hibrid akan mengukur kemampuan pembelajaran jaringan saraf kuantum dalam lanskap pengoptimalan non-cembung dari perspektif kesalahan generalisasi dan kesalahan estimasi akibat pengukuran kuantum. Implementasi kesatuan yang dapat dilatih pada sirkuit kuantum berparameter kemudian dapat dianalisis dengan properti tanpa penyesalan dalam kesalahan yang dapat ditoleransi dari data yang dihasilkan. Kemudian tugas pemrosesan informasi kuantum dapat dirumuskan ulang sebagai pembelajaran diskriminatif, pembelajaran generatif, dan masalah pembelajaran permusuhan dengan kriteria keterpisahan untuk membedakan keterjeratan keadaan kuantum tertentu yang dicirikan dalam data terstruktur dan efisiensi komputasi.

Generalisasi kuantum jaringan permusuhan generatif (GAN) ke rezim mekanika kuantum mencakup struktur pemodelan dalam GAN konvolusional, kondisional, dua arah, dan semi-supervisi menggunakan sirkuit kuantum. Pendekatan kuantum terhadap masalah pemodelan generatif dalam pembelajaran mendalam adversarial teoretis permainan kemudian dapat fokus pada algoritma kuantum variasional dengan metode rantai Markov dalam pelatihan adversarial dan permainan Bayesian Stackelberg dalam pembelajaran mendalam adversarial. Kita juga dapat mengkarakterisasi masalah klasifikasi kuantum dengan adanya noise. Kemudian kita dapat mempelajari trade-off antara kemampuan belajar dan ketahanan pembelajaran mesin yang berlawanan.

Algoritma ilmu informasi kuantum yang efisien dengan pengklasifikasi dasar seperti itu akan membantu kita menangani “data besar” di komputer kuantum. Untuk mendapatkan jaminan yang dapat diandalkan pada evaluasi keamanan jaringan saraf kuantum, kita dapat memperoleh kesalahan generalisasi untuk masalah pembelajaran permusuhan kuantum dengan verifikasi formal. Peningkatan kuantum dari prosedur komputasi ini dalam pembelajaran mendalam adversarial dapat menciptakan algoritme pembelajaran mesin adversarial kuantum dengan produk dalam dari data besar, status kuantum peringkat baris yang dipelajari, transformasi kesatuan sebagai masalah penyimpanan-pengambilan dalam penambangan data, dan tomografi status/proses dari tugas pembelajaran pengukuran kuantum.

Pengorbanan desain pembelajaran mendalam berikut ini harus diatasi dalam pekerjaan di masa depan sebagai gagasan tentang manfaat dan biaya yang dikeluarkan untuk melatih ulang model dan menghasilkan penyerang.

- **Pengorbanan penyerang**: biaya adaptasi terhadap pengklasifikasi dan manfaat serangan

- **Pengorbanan bagi pembela**: manfaat dari deteksi serangan yang benar dan kerugian jika terjadi alarm palsu
- **Pertukaran interaksi**: ruang pencarian strategi, fungsi pembayaran, biaya partisipasi dalam permainan, peringkat relatif setiap pemain yang menyimpulkan batasan keputusan, skenario serangan kotak hitam di mana musuh tidak dapat mengamati strategi pengklasifikasi sebelum memilih strateginya
- **Pertukaran keseimbangan**: tujuan pengoptimalan, kinerja kesalahan klasifikasi, dan biaya pelatihan ulang dalam pembelajaran mendalam
- **Pengorbanan utilitas**: kerugian utilitas pemain bertahan dalam permainan lebih rendah dibandingkan utilitas musuh dan pentingnya biaya yang dikeluarkan untuk menghasilkan serangan dan untuk melatih kembali pengklasifikasi
- **Pertukaran diskriminasi**: kemampuan belajar yang lebih tinggi akan mengakibatkan kurangnya ketahanan

Untuk mengkarakterisasi sinyal data permusuhan dalam data eksperimen, kita juga dapat memperkirakan eksponen spektrum Lyapunov dan jaringan penarik dengan model pembelajaran mendalam dan proses analisis prediktif.

Di sini, kita dapat mengeksplorasi teknik pemrosesan sinyal spektral untuk (i) mendeskripsikan dinamika kompleks dalam konsep solusi teoretis permainan sebagai kesalahan pemodelan dalam prediksi multivariat dengan pembelajaran mendalam, (ii) ekstraksi fitur untuk model pembelajaran mesin pada pembangkitan data dasar dan pelatihan model yang memvalidasi, (iii) merekonstruksi persamaan diferensial sebagai model dinamis dari data pelatihan yang tersedia untuk pelatihan adversarial, dan (iv) analisis lanskap energi dengan skema distribusi data dan struktur pengindeksan data untuk data statis dan dinamis yang mengurangi biaya komunikasi dan meningkatkan penyeimbangan beban dalam sistem memori terdistribusi.

Kita perlu membuat kriteria validasi statistik untuk pembelajaran mesin tersebut dengan mengacu pada pengetahuan domain aplikasi. Dalam konteks ini, kita dapat mengeksplorasi metrik evaluasi dalam penambangan data yang diterapkan pada pemodelan data permusuhan dalam aplikasi keamanan siber.

BAB 5
MEKANISME PERTAHANAN MANIPULASI PEMBELAJARAN MESIN

Dalam bab ini kita mengeksplorasi arsitektur jaringan saraf, implementasi, analisis biaya, dan proses pelatihan menggunakan pembelajaran mendalam adversarial teori permainan. Kami juga mendefinisikan batasan utilitas jaringan saraf dalam dalam teori pembelajaran komputasi seperti minimalisasi risiko empiris, kerangka batasan kesalahan, dan pembelajaran tanpa penyesalan. Di sini kerangka batas kesalahan dengan properti tanpa penyesalan untuk pembelajaran online memberikan kesalahan yang dapat ditoleransi dan memperbarui aturan untuk melatih jaringan saraf pada data manipulasi yang dihasilkan. Kemudian tugas pemrosesan informasi dunia maya dapat dirumuskan ulang menjadi pembelajaran diskriminatif, pembelajaran generatif, dan pembelajaran manipulasi sehubungan dengan kriteria keterpisahan yang menjadi ciri kumpulan data terstruktur dan efisiensi komputasi untuk menyusun fungsi kerugian manipulasi.

Teknik pertahanan proaktif untuk setiap contoh manipulasi juga dirangkum untuk membangun lingkungan pertahanan yang mendalam dengan permainan sinyal manipulasi untuk memitigasi serangan siber oleh musuh yang adaptif. Mereka dapat dimasukkan ke dalam paradigma desain-untuk-keamanan untuk hipotesis pembelajaran mesin guna melengkapi paradigma desain-untuk-kinerja klasik guna menghasilkan berbagai tingkat pertahanan terhadap serangan siber dengan mengacu pada tujuan keamanan perspektif sistem pembelajaran. Kami menyajikan sejumlah besar literatur terkait tentang algoritma optimasi komputasi dalam mekanisme pertahanan untuk persyaratan keamanan dalam pembelajaran mendalam manipulasi, teoritis permainan dengan cara yang sangat terorganisir.

Mereka dapat digunakan untuk forensik digital, identifikasi kerentanan, analisis dampak, mitigasi risiko, metrik keamanan siber, pengembangan data dan model, pengujian penetrasi, dan interoperabilitas semantik dalam aplikasi keamanan siber. Kami menunjukkan banyak penerapan, keterbatasan metode saat ini, arah masa depan yang menjanjikan untuk pengembangan penanggulangan pembelajaran mendalam teori permainan, dan evaluasi teknologi secara rinci. Asumsi pembelajaran mendalam adversarial yang diformalkan untuk permainan yang kuat dapat melacak timbulnya permukaan serangan, kapasitas, dan spesifisitas dalam tujuan multitask yang kritis terhadap keselamatan untuk ketahanan adversarial dalam desain fungsi kerugian dalam mekanisme perlindungan infrastruktur penting.

Pengambilan keputusan algoritmik yang dihasilkan memperjelas kemampuan sistem pembelajaran dalam hal efisiensi, objektivitas, dan kontrol kepada audiens tertentu untuk memungkinkan akurasi, keadilan, akuntabilitas, ketersediaan, integritas, kerahasiaan, stabilitas, keandalan, keamanan, pemeliharaan, dan transparansi. Dalam konteks ini, pembelajaran mendalam manipulasi dapat dirumuskan berdasarkan prinsip AI komposit dan

AI yang dapat dijelaskan yang meningkatkan efisiensi pembelajaran yang diawasi, pembelajaran penguatan, pembelajaran generatif, dan pembelajaran teori permainan dengan representasi pengetahuan untuk penalaran tentang privasi, kepercayaan, dan keamanan. metrik pengoptimalan.

Brendel dkk. mengkategorikan model ancaman yang menghasilkan gangguan manipulasi menjadi (i) serangan berbasis gradien yang mengandalkan informasi model terperinci, (ii) serangan berbasis skor yang mengandalkan skor keyakinan seperti probabilitas yang dikondisikan kelas, (iii) serangan berbasis transfer yang mengandalkan model pengganti untuk model target, dan (iv) usulan serangan berbasis keputusan yang mengandalkan informasi tentang keputusan model akhir. Serangan berbasis keputusan yang diusulkan disebut serangan batas dan diterapkan pada model kotak hitam untuk model target.

5.1 MENGAMANKAN PENGKLASIFIKASI TERHADAP SERANGAN FITUR

Li dkk. mendemonstrasikan keterbatasan pengurangan fitur dalam pengaturan manipulasi dengan musuh yang didorong oleh tujuan. Setiap musuh seharusnya mampu melakukan substitusi pada fitur serupa dalam serangan substitusi silang fitur. Musuh juga diasumsikan mampu mengklasifikasikan kueri berdasarkan anggaran kueri tetap dan anggaran biaya. Model penghindaran dengan pengatur renggang disajikan dalam suasana manipulasi. Membangun pengklasifikasi pada kelas kesetaraan fitur daripada ruang fitur diusulkan sebagai solusi untuk meningkatkan ketahanan pengklasifikasi terhadap model penghindaran. Solusi lain mengusulkan permainan interaksi Stackelberg dua tingkat antara pengklasifikasi dan kumpulan musuh. Permainan Stackelberg diselesaikan dengan pemrograman linier bilangan bulat campuran dengan pembangkitan kendala. Sasaran musuh disimpulkan dari pembangkitan kendala (anggaran kueri dan anggaran biaya) yang menyatu dengan optimal lokal pada data pelatihan.

Globerson dkk. menganalisis ketahanan pengklasifikasi dengan formulasi teori permainan. Untuk pengklasifikasi yang dilatih pada beberapa fitur dengan tingkat kepentingan yang berbeda-beda, fitur tunggal apa pun tidak diberi terlalu banyak bobot selama pengujian. Musuh seharusnya dapat menghapus fitur dalam data pengujian yang ada dalam data pelatihan. Kemudian pengklasifikasi dibuat yang optimal dalam skenario penghapusan fitur terburuk. Skenario seperti ini dirumuskan sebagai solusi permainan dua pemain antara pengklasifikasi dan penghapus fitur dengan tujuan minmax.

Pengklasifikasi memilih tindakan yang memberikan parameter pengklasifikasi yang kuat. Penghapus fitur memilih untuk menghapus fitur yang paling berbahaya bagi kinerja pengklasifikasi. Struktur ketidakpastian dalam konvergensi game berkaitan dengan keberadaan dan ketidakberadaan suatu fitur. Mesin vektor pendukung dengan kehilangan engsel yang teratur dan batasan linier dianggap sebagai tujuan pelatihan untuk pengklasifikasi. Game menghapus fitur yang menyebabkan penurunan maksimum pada hilangnya pengklasifikasi. Permainan kooperatif dengan tujuan nilai Shapley yang mengukur perubahan kinerja setelah menghapus sebuah fitur merupakan alternatif dari tujuan minmax yang diusulkan dengan menghapus beberapa fitur secara bersamaan.

Mengingat skenario serangan penghindaran, Zhang dkk. menyelidiki dampak pengurangan fitur pada keamanan pengklasifikasi, jika pemilihan fitur yang sadar akan musuh tidak ada dalam pelatihan pengklasifikasi. Kumpulan fitur yang lebih kecil terbukti secara signifikan memperburuk kinerja pengklasifikasi yang sedang diserang yang biasanya tidak terjadi pada pengklasifikasi yang tidak diserang. Model keamanan pengklasifikasi dinyatakan sebagai pengatur yang akan dioptimalkan dan diperkirakan bersama dengan kemampuan generalisasi pengklasifikasi selama proses pemilihan fitur. Ini diimplementasikan sebagai metode pemilihan fitur berbasis wrapper menggunakan seleksi maju dan eliminasi fitur mundur yang cocok untuk pengklasifikasi linier dan non-linier dengan fungsi diskriminan yang dapat dibedakan. Skenario serangan penghindaran dianggap sebagai serangan integritas eksplorasi pada data pengujian yang diumpankan ke pengklasifikasi yang dilatih berdasarkan data pelatihan asli.

Strategi penghindaran optimal dirumuskan sebagai masalah optimasi yang meminimalkan jarak antara contoh adversarial dan data pelatihan sehingga fungsi diskriminan pengklasifikasi salah mengklasifikasikan contoh adversarial. Pendekatan pemilihan fitur yang sadar akan musuh tidak hanya memaksimalkan kemampuan generalisasi pengklasifikasi tetapi juga keamanan pengklasifikasi terhadap serangan penghindaran. Di sini keamanan pengklasifikasi diberi bobot berdasarkan batasan dan parameter spesifik aplikasi, sedangkan kemampuan generalisasi pengklasifikasi diperkirakan berdasarkan fungsi diskriminan yang bergantung pada aplikasi dan ukuran kinerja. Alih-alih mencari titik penghindaran terbaik dengan menanyakan pengklasifikasi dengan sampel kandidat dari pendekatan pencarian kotak hitam, algoritma adversarial yang efisien secara komputasi dirancang untuk mengeksploitasi pengetahuan musuh tentang fungsi tujuan pengklasifikasi yang ditargetkan.

Mereka ditentukan oleh pilihan fungsi jarak antara contoh manipulasi dan data pelatihan serta representasi fitur dalam algoritma klasifikasi. Prosedur penurunan gradien menemukan langkah-langkah gradien yang mengurangi jarak antara data manipulasi dan data pelatihan sambil memproyeksikan titik saat ini ke domain contoh manipulasi yang layak segera setelah fungsi diskriminan salah mengklasifikasikannya. Titik serangan awal dalam penurunan gradien diatur ke sampel terdekat yang diklasifikasikan sebagai sah atau diklasifikasikan sebagai berbahaya.

Kinerja pengklasifikasi sejati menurun dengan baik terhadap serangan dengan peningkatan kekuatan serangan yang ditentukan oleh batas atas pada jumlah maksimum modifikasi manipulasi dan batas bawah pada kepercayaan pengklasifikasi yang menyesatkan. Pengklasifikasi yang terdegradasi dengan baik diharapkan menjadi yang paling aman setelah pelatihan ulang pada data pelatihan serta contoh-contoh yang berlawanan. Eksperimen memvalidasi uji-t siswa dan akurasi klasifikasi distribusi bobot fitur dilakukan pada korpus email TREC 2007 yang terdiri dari email sah dan email spam. Batasan khusus aplikasi pada distribusi data mempersulit musuh untuk meniru nilai fitur dari kelas yang sah, yang pada akhirnya menyebabkan rendahnya kemungkinan untuk menghindari deteksi.

5.2 TUGAS KLASIFIKASI PEMBELAJARAN MESIN DENGAN REGULARIZER

Demontis dkk. menganalisis serangan penghindaran pengklasifikasi linier dalam kerangka optimasi yang kuat. Hubungan antara ketersebaran bobot fitur dan pertahanan pengklasifikasi linier diselidiki untuk mengusulkan pengatur. Pengklasifikasi linier dipilih dalam algoritma pembelajaran adversarial karena keputusannya yang dapat ditafsirkan diperoleh dari rendahnya penyimpanan, waktu pemrosesan, dan konsumsi daya dalam sistem seluler dan tertanam. Musuh seharusnya memiliki pengetahuan lengkap tentang data pelatihan pengklasifikasi target, kumpulan fitur, dan algoritma klasifikasi. Kemampuan Musuh dalam memodifikasi data diberikan sebagai batasan data yang bergantung pada aplikasi.

Biasanya, batasan data tersebut didefinisikan sebagai 41 dan 41 norma pada sejumlah fitur yang dimodifikasi yang masing-masing disebut serangan jarang dan padat. Strategi serangan Adversary dirumuskan sebagai masalah optimasi yang meminimalkan fungsi diskriminan pengklasifikasi target untuk data yang memiliki batasan jarak antara contoh adversarial dan data asli. Dengan gagasan untuk menemukan bobot yang jarang dan seragam, kombinasi cembung linier dari 41 dan 4 norma diusulkan sebagai pengatur ketahanan strategi serangan musuh. Perilaku regularisasi tersebut terhadap serangan penghindaran pada pengklasifikasi mesin vektor dukungan dengan kehilangan engsel kemudian diselidiki dalam aplikasi klasifikasi untuk klasifikasi digit tulisan tangan, pemfilteran spam, dan deteksi malware. Pengukuran kinerja dalam pengaturan manipulasi dilakukan dengan area di bawah kurva ROC dikombinasikan dengan ketersebaran dan langkah-langkah keamanan yang diusulkan pada distribusi bobot pengklasifikasi.

Krause dkk. menyajikan fungsi tujuan teori informasi untuk melatih pengklasifikasi probabilistik diskriminatif yang disebut Regularized Information Maximization (RIM). RIM diterapkan sebagai kerangka pengelompokan yang mengakomodasi fungsi kemungkinan yang berbeda, menyeimbangkan pemisahan kelas, dan menggabungkan label parsial untuk pembelajaran semi-supervisi. Teknik pengelompokan diskriminatif tersebut mewakili batas-batas antara kategori pengelompokan yang tersedia dalam aplikasi pengelompokan di dunia nyata. Ini mencakup teknik seperti partisi grafik spektral, pengelompokan margin maksimum, dan model gas saraf.

Di sini pembelajaran masalah pengelompokan tanpa pengawasan diformalkan sebagai model probabilistik bersyarat yang cocok untuk pengelompokan diskriminatif kelas jamak. Fungsi tujuan kemudian memaksimalkan informasi timbal balik antara distribusi data empiris pada input dan distribusi label terinduksi dari pemilihan model. Ini dibangun untuk memenuhi sifat matematika untuk optimasi seperti batas keputusan tidak boleh ditempatkan di ruang input yang padat dengan titik data dan konfigurasi pengelompokan di mana label kategori didistribusikan secara merata di seluruh kelas lebih disukai. Selanjutnya istilah pengatur diperkenalkan untuk menghukum model kondisional dengan batasan keputusan yang kompleks dalam pemilihan model. Hal ini bergantung pada pilihan spesifik dari distribusi probabilitas bersyarat yang diestimasi. Dalam masalah klasifikasi kelas jamak, keyakinan sebelumnya tentang proporsi label kelas yang tidak seragam dikodekan sebagai istilah entropi relatif dalam fungsi tujuan non-cembung RIM.

Xu dkk. membuat mesin vektor dukungan (SVM) yang diatur dalam formulasi pengoptimalan yang kuat berdasarkan kumpulan ketidakpastian. SVM semacam itu memiliki perlindungan terhadap kebisingan dan overfitting. Mereka meminimalkan kombinasi kesalahan pelatihan dan jangka waktu regularisasi. Istilah regularisasi biasanya berupa norma tensor. Ini membatasi kompleksitas kelas fungsi pengklasifikasi untuk mendukung kinerja generalisasi. Ini menganggap sampel data pengujian sebagai salinan sampel data pelatihan yang terganggu.

Oleh karena itu, membatasi gangguan tersebut akan mengurangi kesenjangan antara kesalahan klasifikasi. Pendekatan minimalisasi risiko struktural adalah teknik regularisasi yang meminimalkan batasan kesalahan generalisasi berdasarkan kesalahan pelatihan dan istilah kompleksitas. SVM tangguh yang diusulkan melakukan optimasi minmax atas semua kemungkinan gangguan antara sampel data pelatihan dan pengujian. Stabilitas SVM terhadap gangguan tertentu yang dapat diperkirakan merupakan gagasan ketahanan terkait yang juga dipelajari. Kerugian pelatihan ditambah penalti regularisasi adalah kerugian yang diatur untuk melatih SVM yang kuat.

Yan dkk. mengusulkan jaringan maksimalisasi margin adversarial (AMM) yang memiliki regularisasi pembelajaran adversarial berbasis perturbasi adversarial. Formulasi gangguan yang dapat dibedakan disebarkan kembali melalui jaring dalam yang diatur. Pengklasifikasi margin maksimum seperti itu cenderung memiliki kinerja generalisasi yang lebih baik karena kekompakan antar kelas dan kemampuan diskriminasi antar kelas. Mekanisme pertahanan manipulasi yang diusulkan dapat digeneralisasi menjadi pengklasifikasi multi-label selama label target dipilih dengan tepat untuk gangguan manipulasi.

Zhong dkk. menyematkan istilah regularisasi berbasis margin ke dalam tujuan klasifikasi jaringan saraf dalam. Istilah regularisasi memiliki dua langkah optimasi untuk menemukan potensi gangguan secara berulang. Margin yang besar dalam klasifikasi adversarial menjamin jarak antar kelas dan kelancaran antar kelas dalam ruang penyematan untuk meningkatkan ketahanan jaring dalam. Fungsi kerugian lintas entropi dioptimalkan bersama dengan batasan jarak margin besar yang bertindak sebagai istilah regularisasi. Ketahanan pengklasifikasi diuji dalam kondisi manipulasi fitur dan manipulasi label.

Alabdulmohsin dkk. membahas serangan rekayasa balik terhadap pengklasifikasi dengan batasan keputusan tetap. Kemudian pengacakan dalam klasifikasi akibat pemrograman semidefinite dalam suatu distribusi pengklasifikasi dirumuskan untuk memitigasi risiko manipulasi dan memberikan prediksi yang andal dengan probabilitas yang tinggi. Penulis menyelidiki trade-off antara akurasi prediksi dan varians distribusi pengklasifikasi. Serangan rekayasa balik yang diusulkan diklasifikasikan dalam skenario serangan eksplorasi di mana musuh memanipulasi distribusi data pengujian. Sistem klasifikasi yang diusulkan berupaya membuat prediksi yang andal sambil mengungkapkan informasi sesedikit mungkin tentang batas-batas keputusan.

Masalah pembelajaran dengan distribusi pengklasifikasi dirumuskan sebagai masalah optimasi cembung. Pertahanan sistem klasifikasi dibandingkan dengan klasifikasi adversarial, koreksi matriks kernel, pembelajaran ensembel, pembelajaran multi-contoh, dan mekanisme

pembelajaran adversarial teori permainan. Di sini strategi pertahanan eksplorasi dikatakan menyebabkan disinformasi tentang pilihan data pelatihan, fitur, fungsi biaya, dan algoritma pembelajaran. Strategi pertahanan eksplorasi lainnya adalah meningkatkan kompleksitas ruang hipotesis bagi musuh tanpa menyebabkan overfitting pada pengklasifikasi. Dalam kasus seperti ini, strategi pengacakan akan memperkirakan probabilitas pemilihan label kelas alih-alih memprediksinya sebagai label biner.

Tujuan dari pengacakan yang berhasil adalah untuk meningkatkan upaya rekayasa balik musuh tanpa meningkatkan tingkat kesalahan prediksi pengklasifikasi. Selanjutnya algoritma pembelajaran aktif diusulkan kepada musuh untuk membuat kueri target pada pengklasifikasi. Di sini strategi pemilihan kueri didasarkan pada pengambilan sampel acak, pengambilan sampel selektif, dan pengambilan sampel ketidakpastian di mana musuh mengetahui bahwa pembela menggunakan pengklasifikasi acak. Setelah pembelajaran diselesaikan oleh pembela HAM, pembela HAM dapat memitigasi risiko manipulasi akibat serangan rekayasa balik dengan memilih pengklasifikasi secara acak dari distribusi pengklasifikasi untuk setiap kueri yang diamati dari sisi musuh.

Pengklasifikasi linier digunakan untuk membangun ansambel distribusi pengklasifikasi. Dalam evaluasi eksperimental, kurva tradeoff akurasi-varians dibuat untuk menganalisis titik optimalitas Pareto dari sistem klasifikasi. Setiap titik Pareto-optimal merupakan strategi yang baik untuk mempertahankan sistem klasifikasi. Dengan menggambar pengklasifikasi secara acak dari distribusi dengan varian besar, kompleksitas komputasi musuh untuk melaksanakan serangan rekayasa balik yang diusulkan meningkat secara signifikan dengan sedikit peningkatan biaya komputasi untuk sistem pembelajaran. Model klasifikasi seperti ini cocok untuk diterapkan pada aplikasi yang sensitif terhadap keamanan seperti pemfilteran spam, deteksi intrusi, dan deteksi penipuan.

Zhang dkk. mengusulkan pelatihan mesin adversarial untuk menggunakan data adversarial paling sedikit untuk memperbarui model pembelajaran. Tujuan pembelajaran manipulasi memperoleh batas atas risiko manipulasi. Risiko pelatihan mesin adversarial, melatih jaringan neural dalam menggunakan data adversarial yang diprediksi salah untuk meminimalkan kerugian dan memprediksi data adversarial dengan tepat untuk memaksimalkan kerugian. Penurunan gradien yang diproyeksikan (PGD) dengan penghentian awal digunakan untuk membuat data manipulasi dalam proses pelatihan. Pembelajaran kurikulum digunakan dalam strategi pelatihan manipulasi yang diusulkan untuk meningkatkan ketahanan jaringan saraf dalam. Artinya, jaringan saraf dalam awalnya belajar dari data manipulasi yang lebih ringan dan kemudian secara bertahap beradaptasi dengan data manipulasi yang lebih kuat. Sinha dkk. melakukan analisis teoritis tentang optimasi distribusi yang kuat untuk pelatihan manipulasi. Prosedur pelatihan menambah parameter pemodelan dengan gangguan terburuk pada data pelatihan. Ini menyatu dengan model pembelajaran yang mencapai ketahanan dengan biaya statistik atau komputasi yang kecil dibandingkan dengan minimalisasi risiko empiris.

Tsipras dkk. mempelajari trade-off antara kinerja generalisasi standar dan ketahanan manipulasi terhadap contoh-contoh manipulasi dalam pembelajaran mesin. Argumen yang

disajikan adalah bahwa pengklasifikasi kuat mempelajari representasi yang berbeda secara statistik dibandingkan pengklasifikasi standar. Tujuan dari manipulasi pembelajaran mesin didefinisikan sebagai model pelatihan dengan perkiraan kerugian adversarial yang rendah dengan adanya gangguan input kasus terburuk sebagai contoh adversarial. Pengklasifikasi standar terbukti memanfaatkan fitur-fitur yang berkorelasi lemah dengan label kelas untuk mencapai akurasi standar. Sebaliknya, manipulasi dapat mensimulasikan distribusi fitur-fitur yang berkorelasi lemah seolah-olah fitur-fitur tersebut termasuk dalam kelas yang salah.

Oleh karena itu, pengklasifikasi standar apa pun yang bertujuan untuk mendapatkan akurasi tinggi harus bergantung pada fitur tidak kuat yang dapat dimanipulasi secara sewenang-wenang. Selain itu, trade-off antara akurasi standar dan akurasi adversarial melekat pada distribusi data yang mendasarinya dan bukan karena kurangnya sampel untuk pelatihan. Contoh-contoh manipulasi yang dihasilkan dari gangguan fitur-fitur yang tidak kuat tersebut akan ditransfer ke semua pengklasifikasi yang mengandalkan fitur-fitur yang berkorelasi lemah dengan label kelas yang benar. Dalam data pelatihan yang terbatas, fitur rapuh seperti itu bahkan dapat muncul karena noise. Oleh karena itu, gangguan manipulasi dapat diartikan sebagai sifat invarian yang dapat dipenuhi oleh model yang kuat. Pelatihan yang kuat yang menghasilkan kerugian kecil untuk semua gangguan dapat dipandang sebagai metode untuk menanamkan invarian tertentu dalam model klasifikasi standar.

Dalam konteks ini, penulis mengamati bahwa gradien untuk jaringan saraf yang dilatih secara musuh selaras dengan fitur gambar masukan yang relevan secara persepsi. Jadi kita dapat menafsirkan gangguan manipulasi sebagai menghasilkan karakteristik yang menonjol dari sampel yang termasuk dalam kelas target yang diinterpolasi. Penjelasan seperti itu tidak dapat diberikan dalam model standar di mana contoh-contoh manipulasi muncul sebagai varian berisik dari gambar masukan. Kelas target yang diinterpolasi dapat direpresentasikan dengan model generatif yang mendalam seperti jaringan manipulasi generatif dan autoencoder variasional yang melibatkan manipulasi ke dalam representasi yang dipelajari. Lanskap kerugian model pembelajaran yang kuat kemudian dapat digunakan untuk melakukan interpolasi antar kelas dengan lancar. Mempelajari asumsi generatif dalam data memungkinkan kami memberikan batas atas ketahanan pengklasifikasi yang mampu memperhitungkan kompleksitas sampel pembelajaran yang kuat.

5.3 PEMBELAJARAN MESIN PENGUATAN

Pembelajaran mesin penguatan adalah studi tentang agen cerdas dan tindakan mereka dalam lingkungan simulasi sehingga gagasan tentang imbalan kumulatif dimaksimalkan dalam interaksi antara agen dan lingkungan. Alih-alih label input/output yang diperlukan dalam pembelajaran mesin yang diawasi, fokus pembelajaran penguatan adalah menemukan keseimbangan antara eksplorasi dan eksploitasi pola. Pembelajaran penguatan dapat diartikan sebagai metode berbasis pengambilan sampel untuk menyelesaikan masalah pengendalian yang optimal.

Tujuan pembelajaran penguatan adalah mempelajari kebijakan yang memaksimalkan imbalan kumulatif yang diharapkan dan meminimalkan penyesalan jangka panjang. Agen

cerdas dalam pembelajaran penguatan harus memilih tindakan secara acak tanpa mengacu pada perkiraan distribusi probabilitas. Tugas pembelajaran penguatan asosiatif menggabungkan pembelajaran yang diawasi dengan pembelajaran penguatan. Dalam pemodelan teori permainan, pembelajaran penguatan dapat digunakan untuk menghasilkan estimasi kesalahan pada optimasi dengan mengacu pada rasionalitas terbatas.

Chen dkk. meninjau taksonomi serangan manipulasi pada pembelajaran penguatan. Contoh manipulasi diklasifikasikan menjadi contoh manipulasi implisit yang menambahkan manipulasi yang tidak terlihat untuk menyesatkan pelajar dan contoh manipulasi dominan yang menambahkan gangguan dunia fisik untuk mengubah informasi lokal yang tersedia untuk pembelajaran penguatan. Skenario serangan adversarial diklasifikasikan menjadi serangan kesalahan klasifikasi untuk menargetkan jaringan saraf yang melakukan pembelajaran penguatan dan serangan bertarget untuk menargetkan label kelas tertentu dalam pelatihan yang salah diklasifikasikan ke dalam label kelas target yang dipilih oleh musuh.

Model pembelajaran yang dilatih menurut kebijakan pembelajaran penguatan disebut agen target. Q-Learning adalah algoritma pelatihan populer untuk pembelajaran penguatan. Ini mengusulkan pembaruan pada nilai Q yang mewakili imbalan kumulatif dari agen target. Melalui proses pembelajaran berulang, agen target memaksimalkan nilai Q dengan menemukan jalur terbaik menuju tujuan. Hal ini dapat diwakili oleh fungsi utilitas yang mengevaluasi kekuatan dan kelemahan tindakan dalam keadaan tertentu. Deep Q-Network adalah peningkatan pembelajaran mendalam pada Q-Learning. Hal ini memunculkan pembelajaran penguatan mendalam dengan fungsi kerugian jaringan pembelajaran mendalam yang mendefinisikan utilitas nilai-Q.

Algoritme A3C (asynchronous advantage actor-critic) menggunakan kerangka aktor-kritikus untuk meningkatkan proses pelatihan dalam pembelajaran penguatan mendalam. Trust Region Policy Optimization (TRPO) mampu mengendalikan perubahan penguatan kebijakan pembelajaran dari perbedaan teori informasi KL kebijakan lama dan kebijakan baru. Tinjauan literatur selanjutnya oleh Chen et al. menunjukkan bahwa metode tanda gradien cepat (FGSM) dapat disesuaikan dengan sistem pembelajaran penguatan dan contoh manipulasi dapat dibuat untuk jalur pembelajaran Q dari gradien nilai Q maksimum untuk setiap titik di jalur tersebut. Serangan induksi kebijakan dirangkum untuk Deep Q-Networks.

Mekanisme pertahanan manipulasi diusulkan karena varian pelatihan manipulasi dan regularisasi tujuan pembelajaran dalam fungsi kerugian manipulasi untuk pembelajaran penguatan mendalam. Dalam situasi serangan seperti itu, model ancaman blackbox yang lengkap cukup jarang terjadi. Variasi istilah pelatihan dan regularisasi manipulasi dalam fungsi tujuan, modifikasi struktur jaringan seperti distilasi defensif, dan pemodelan generatif mendalam yang menghasilkan contoh manipulasi adalah mekanisme pertahanan yang paling umum. Domain aplikasi untuk pembelajaran mesin manipulasi tersebut mencakup pemahaman bahasa alami, pemahaman gambar, pengenalan suara, mengemudi otonom, navigasi visual berbasis target, permainan game, sistem perdagangan, sistem pemberi rekomendasi, sistem dialog, manajemen inventaris, dan perencanaan jalur otomatis. Survei

konsep solusi teoretis permainan dalam pembelajaran penguatan mendalam multi-agen diberikan oleh Lu et al.

Dai dkk. fokus pada serangan manipulasi yang mengubah struktur kombinatorial data dalam domain aplikasi yang melibatkan struktur data grafik. Metode serangan berbasis pembelajaran penguatan diusulkan untuk menyusun kebijakan serangan dari umpan balik prediksi dari pengklasifikasi target. Pengklasifikasi target dibuat dengan model jaringan neural grafik yang melakukan tugas klasifikasi tingkat grafik dan tingkat simpul. Kelompok model pembelajaran terawasi yang dianalisis memiliki penerapan dalam tugas transduktif dan tugas induktif. Tidak seperti serangan adversarial pada gambar yang merupakan kumpulan data berkelanjutan, serangan adversarial pada grafik harus dilakukan pada kumpulan data terpisah.

Manipulasi tersebut dilakukan dengan menambahkan atau menghilangkan sisi-sisi grafik secara berurutan. Kompleksitas waktu kuadrat dari ruang tindakan pada node grafik diatasi dengan teknik berbasis dekomposisi grafik. Model ancaman diklasifikasikan menjadi (i) serangan kotak putih (white-box) dimana musuh memiliki akses ke internal pengklasifikasi target termasuk label prediksi, informasi gradien, dll., (ii) serangan kotak hitam (blackbox) dimana hanya prediksi dari pengklasifikasi target yang tersedia untuk digunakan. musuh, dan (iii) membatasi serangan blackbox dimana musuh dapat melakukan query blackbox pada beberapa sampel untuk dapat membuat manipulasi pada sampel yang tersisa. Serangan yang tidak ditargetkan adalah fokus dari manipulasi. Penelitian ini juga dapat diperluas ke serangan yang ditargetkan.

Fungsi kerugian lintas entropi digunakan dalam pelatihan pengklasifikasi. Penyematan fitur tingkat grafik dan tingkat simpul digunakan untuk melatih jaringan saraf grafik. Indikator kesetaraan grafik diusulkan untuk memenuhi syarat semantik klasifikasi sebelum dan sesudah manipulasi. Fungsi penghargaan diusulkan untuk musuh yang bertindak sebagai agen pembelajaran penguatan. Algoritme Q-learning kemudian mempelajari proses keputusan Markov (MDP) untuk memecahkan masalah optimasi diskrit dengan horizon terbatas. Setiap sampel manipulasi yang dihasilkan mendefinisikan MDP tersebut. Untuk mempelajari musuh yang dapat digeneralisasikan, tujuan pembelajaran fungsi Q dalam pembelajaran Q digeneralisasikan untuk mentransfer seluruh sampel musuh dan MDP terkaitnya.

Selanjutnya, metode serangan blackbox diusulkan dengan algoritma genetika untuk skenario optimasi orde nol. Tujuan optimasi dalam skenario optimasi orde nol diselesaikan dengan algoritma optimasi bebas turunan. Metode beda hingga pada nilai fungsi digunakan untuk memperkirakan gradien yang dibentuk oleh turunan terarah dari fungsi kerugian yang ditargetkan. Kriteria konvergensi untuk estimasi tersebut bergantung pada kompleksitas iterasi optimasi dan kompleksitas kueri evaluasi fungsi. Versi algoritma optimasi bebas turunan yang tidak dibatasi seperti metode pengali arah bolak-balik (ADMM) meminimalkan fungsi kerugian rata-rata empiris yang non-cembung.

Versi dua tingkat dari algoritme pengoptimalan bebas turunan merumuskan tujuan teoretis permainan yang biasanya merupakan fungsi minmax dalam serangan kotak hitam. Hal ini diselesaikan dengan algoritma seperti penurunan koordinat stokastik orde nol.

Kompleksitas komputasi dari algoritme ini diatasi dengan teknik seperti reduksi dimensi dan pengambilan sampel kepentingan. Algoritme untuk pembelajaran mendalam adversarial teoretis permainan juga sebanding dengan formulasi fungsi tujuan yang mempelajari ketahanan adversarial dari pengklasifikasi yang sensitif terhadap biaya. Kerangka minimalisasi penyesalan dapat digunakan untuk mengatasi masalah kompleksitas komputasi yang dihadapi oleh musuh teoritis permainan tersebut.

Mandlekar dkk. mensintesis serangan kotak putih dalam kebijakan pembelajaran penguatan mendalam. Behzadan dkk. mendemonstrasikan contoh manipulasi yang dapat ditransfer ke berbagai Deep Q-Networks. Ciri-ciri spatiotemporal dari proses pelatihan diduga memberikan mekanisme pertahanan terhadap contoh-contoh manipulasi tersebut. Kos dkk. membuat serangan keracunan seiring waktu dalam pembelajaran penguatan mendalam. Contoh manipulasi dalam pengaturan klasifikasi gambar dibandingkan dengan contoh manipulasi dalam pengaturan pembelajaran penguatan.

Ketahanan kebijakan agen pembelajaran melalui pelatihan ulang juga diselidiki. Ilyas dkk. meningkatkan skenario serangan blackbox dengan estimasi gradien berbasis optimasi bandit. Pinto dkk. melatih agen pembelajaran penguatan di hadapan musuh yang mengganggu stabilitas. Musuh menerapkan perbedaan dalam kondisi pelatihan dan pengujian sebagai kekuatan pengganggu dalam pembelajaran penguatan. Lintasan pembelajaran kebijakan kemudian dirumuskan sebagai solusi terhadap permainan Markov yang melibatkan dua pemain (zero-sum Markov game). Li dkk. membahas kendala operasional dalam penghindaran kebijakan keamanan yang bersifat manipulasi. Tugas klasifikasi adversarial dipisahkan menjadi tugas belajar memprediksi preferensi serangan dan tugas mengoptimalkan kebijakan operasional yang secara eksplisit mematuhi batasan operasional pada prediktor. Kemudian strategi respons terbaik pihak lawan dihitung sebagai keputusan operasional yang diacak.

Juni dkk. mengusulkan protokol serangan manipulasi hadiah dalam pembelajaran online dengan umpan balik terbatas. Tujuan manipulasi adalah untuk mendorong atau menghalangi tindakan yang dipilih oleh algoritma bandit kontekstual stokastik. Ma dkk. mengusulkan serangan peracunan data untuk membajak perilaku bandit kontekstual dalam sistem pemberi rekomendasi online. Data adversarial ditemukan dengan menyelesaikan program kuadrat dengan batasan linier. Lin dkk. mengusulkan musuh yang memikat agen melalui serangkaian tindakan yang diinginkan ke negara target yang ditentukan. Model generatif digunakan untuk merencanakan dan memprediksi keadaan agen di masa depan.

Ho dkk. mengusulkan kerangka pembelajaran generatif untuk mempelajari kebijakan pembelajaran imitasi dari lintasan ahli. Kebijakan tersebut dipelajari dengan algoritma generatif yang mengabaikan pembelajaran fungsi biaya entropi kausal maksimum dalam pembelajaran penguatan terbalik. Goyal dkk. melatih generator dalam jaringan manipulasi generatif dengan tujuan perbedaan temporal (TD) daripada gradien diskriminator. Pfau dkk. memandang jaringan manipulasi generatif sebagai metode aktor-kritik di mana aktor tidak dapat mempengaruhi imbalannya. Finn dkk. merumuskan kembali permainan minmax dalam model generatif yang mendalam sebagai masalah optimasi dua tingkat.

Bowling dkk. melaporkan analisis teori permainan stokastik dalam pembelajaran penguatan multi-agen. Meskipun tindakan agen cerdas tunggal dapat dimodelkan sebagai proses keputusan Markov yang stasioner, lingkungan multi-agen harus memodelkan distribusi data non-stasioner di antara beberapa agen yang berinteraksi. Di sini permainan stokastik akan menjadi perpanjangan alami dari proses pengambilan keputusan Markov yang mencakup banyak agen. Permainan stokastik seperti itu pada gilirannya akan menjadi perpanjangan dari permainan matriks. Strategi stasioner dapat dievaluasi dalam permainan matriks hanya jika strategi pemain lain diketahui sebelumnya. Jika tidak, permainan matriks akan melibatkan lingkungan non-stasioner.

Lebih jauh lagi, permainan semacam itu dapat melibatkan strategi murni atau campuran. Dua jenis permainan matriks yang relevan untuk analisis data non-stasioner adalah permainan matriks kolaboratif dan kompetitif yang dikategorikan menurut definisi fungsi imbalannya. Permainan zero-sum dan permainan jumlah umum adalah murni permainan kompetitif. Solusi mereka adalah nilai yang diharapkan dari fungsi pembayaran yang ditemukan masing-masing dengan pemrograman linier dan pemrograman kuadratik.

Pembelajaran penguatan multi-agen mempelajari kebijakan stokastik yang memetakan keadaan beberapa agen saat ini ke distribusi probabilitas atas tindakan mereka. Kebijakan stokastik dapat dianalisis dengan permainan matriks yang diperluas ke beberapa negara bagian yang melibatkan permainan stokastik. Setiap negara bagian dalam permainan stokastik dapat dipahami sebagai permainan matriks yang dimainkan dengan pembayaran bersama dari beberapa agen yang bertransisi antar negara bagian. Strategi keseimbangan cenderung memecahkan masalah komputasi kompleks yang memerlukan teknik pengacakan, generalisasi, dan perkiraan dalam pembelajaran mesin adversarial.

Lanctot dkk. mengusulkan bahwa pembelajaran penguatan multi-agen (MARL) diperlukan untuk mencapai kecerdasan umum buatan. Penulis menyelidiki jenis MARL yang disebut pembelajaran penguatan independen (InRL) yang membuat agen memperlakukan pengalaman pembelajaran mesin mereka sebagai lingkungan non-stasioner. Algoritme pembelajaran kemudian dirancang untuk respons terbaik teoretis permainan yang dihitung untuk campuran kebijakan yang dihasilkan dalam pembelajaran penguatan mendalam. Dengan demikian pemodelan teori permainan digunakan dalam pemilihan kebijakan pembelajaran penguatan.

Analisis algoritmik secara empiris dibandingkan dengan algoritme terkait dalam literatur seperti respons terbaik yang diulang, ramalan ganda, dan permainan fiktif. Ukuran kinerja pembelajaran mesin yang disebut korelasi kebijakan bersama diusulkan untuk mengurangi overfitting dalam generalisasi InRL mulai dari pelatihan hingga eksekusi. Formulasi pembelajaran penguatan InRL memperlakukan setiap agen pelajar sebagai tidak menyadari agen yang tersisa untuk dapat memperlakukan semua interaksinya sebagai milik distribusi data lokal di lingkungan non-stasioner. Lingkungan lokal seperti itu menyebabkan kondisi non-stasioner dan non-Markovian dalam kriteria konvergensi fungsi kerugian adversarial yang diturunkan untuk beberapa algoritma komputasi.

Kemudian kebijakan pembelajaran penguatan untuk InRL dapat disesuaikan dengan lingkungan non-stasioner yang diwakili dalam kebijakan agen yang tersisa sehingga mengakibatkan hilangnya kinerja generalisasi. Reaksi dinamis terhadap perilaku agen dibahas dalam InRL dengan kemampuan observasi parsial dalam pengaturan multi-agen. Untuk menghadapi dinamika pengambilan sampel seperti itu, peneliti harus sering menggunakan perkiraan dalam algoritma pembelajaran yang dihadapkan pada perhitungan yang sulit dilakukan. Lanctot dkk. menggunakan pemodelan teori permainan empiris atas distribusi meta-strategi untuk menghitung respons terbaik atas distribusi kebijakan dalam pembelajaran penguatan mendalam.

Proses pelatihan dengan tabel hasil yang terpusat dan empiris diasumsikan untuk pelaksanaan kebijakan yang terdistribusi dan terdesentralisasi. Algoritma double oracle yang diusulkan menggunakan jaringan saraf dalam sebagai perkiraan fungsi di seluruh iterasi teoretis permainan yang menghitung matriks hasil pada perkiraan strategi respons terbaik. Matriks korelasi kebijakan bersama dihitung untuk menghindari overfitting dalam proses pembelajaran. Tuyl dkk. menganalisis interaksi multi-agen yang kompleks dengan analisis teoritis permainan empiris. Jumlah sampel data yang diperlukan mendekati permainan yang mendasari konvergensi ke ekuilibrium Nash diperiksa. Setiap agen diperlakukan sebagai pemain dengan matriks pembayaran. Sistem dinamika orde pertama dalam teori permainan evolusi merumuskan meta-permainan interaksi kompleks. Tuyls mensurvei penggunaan teori permainan evolusi dalam pembelajaran penguatan dan sistem multi-agen. Kononen membangun metode pembelajaran dalam permainan Markov untuk pembelajaran penguatan multi-agen. Jenis permainan matematika yang diusulkan adalah permainan matriks.

Konsep kesetimbangan Stackelberg dalam permainan tersebut diselesaikan dengan metode pemrograman matematika dalam permainan Markov. Permainan Markov semacam itu memperluas proses pengambilan keputusan Markov ke pengoptimalan pada permainan berulang multi-negara. Aturan pembaruan untuk mengoptimalkan parameter dalam iterasi Q-learning dalam lingkungan online kemudian disajikan. Nowé menganalisis kebijakan optimal agen yang beroperasi di lingkungan pembelajaran penguatan multi-agen dengan teori permainan. Di sini agen harus menghadapi lingkungan komputasi stokastik non-stasioner yang bervariasi menurut kebijakan agen lainnya. Jadi agen harus menemukan solusi yang baik secara statistik untuk pembelajaran mesin dengan berkoordinasi atau bersaing dengan agen lain. Optima teoritis permainan yang ditemukan pada keseimbangan Nash dalam lingkungan seperti itu dianalisis sehubungan dengan algoritma komputasi untuk permainan tanpa kewarganegaraan dengan automata pembelajaran Q, permainan Markov dengan gradien kebijakan, dan pembelajaran tindakan bersama dalam permainan berulang.

Berbeda dengan permainan zero-sum, permainan tersebut merupakan permainan jumlah umum tanpa batasan khusus pada persaingan antar pemain yang berpartisipasi dalam permainan tersebut. Konsep solusi untuk permainan tersebut ditemukan dengan mempelajari strategi respons terbaik untuk setiap pemain yang memaksimalkan hasil saat ini sehubungan dengan strategi lawan saat ini dalam permainan. Ekuilibrium Nash dan

minimalisasi penyesalan adalah konsep keseimbangan paling populer untuk pembelajaran penguatan teoretis permainan. Ini memiliki aplikasi dalam sistem multi-agen. Penyesalan mengacu pada perbedaan antara imbalan yang diharapkan dan aktual bagi seorang agen. Imbalan yang diharapkan dihitung berdasarkan berbagai strategi yang ditetapkan dalam game baik murni atau campuran dalam ruang pencarian untuk pembelajaran mesin. Imbalan sebenarnya dihitung secara empiris selama pelaksanaan permainan. Akumulasi penyesalan dioptimalkan dalam pendekatan pembelajaran berbasis penyesalan. Algoritma populer yang menggabungkan permainan stokastik dengan pembelajaran penguatan adalah Minimax-Q ,[Nash-Q, Fictitious Self-Play , dan minimalisasi penyesalan kontrafaktual.

Lagu dkk. mengusulkan algoritma pembelajaran imitasi pengaturan aktor-kritikus multi-agen. Dalam pembelajaran imitasi, agen mempelajari perilaku yang diinginkan dengan meniru dan ahli. Pakar kira-kira mengoptimalkan fungsi imbalan yang mendasarinya. Agen peniru mempelajari kebijakan melalui pembelajaran penguatan. Dalam pengaturan multi-agen, fungsi penghargaan optima bergantung pada lingkungan non-stasioner dengan beberapa solusi optimal. Algoritme pembelajaran imitasi dari agen tunggal dapat diperluas ke pengaturan multi-agen dalam kerangka pelatihan manipulasi generatif. Penulis memetakan pembelajaran imitasi pada permainan dua pemain antara generator dan diskriminator. Generator mengontrol kebijakan semua agen terdistribusi. Diskriminator adalah pengklasifikasi untuk setiap agen yang membedakan antara perilaku agen dan pakar. Diskriminator memetakan pasangan tindakan negara ke dalam skor. Diskriminator juga dapat memasukkan informasi sebelumnya tentang agen yang bekerja sama dan bersaing. Untuk memaksimalkan fungsi imbalan manipulasi nya, generator mencoba mengelabui diskriminator dengan lintasan sintetik. Pemodelan entropi maksimum membentuk fungsi kerugian untuk estimasi kemungkinan maksimum dalam pembelajaran imitasi yang diusulkan. Pelatihan manipulasi digunakan untuk menggabungkan pengetahuan sebelumnya tentang pengaturan multi-agen dengan fungsi indikator dalam pengatur hadiah yang ditambah dalam permainan minmax untuk pembelajaran penguatan. Algoritma gradien kebijakan yang disebut Kronecker-Factored Trust Region adalah algoritma optimasi yang menyelesaikan konsep keseimbangan teoritis permainan. Pembelajaran imitasi disebut juga pembelajaran penguatan terbalik (IRL).

Multiarmed bandits adalah versi sederhana dari pembelajaran penguatan yang bisa mendapatkan keuntungan dari pelatihan manipulasi. Algoritme multiarmed bandit mengeluarkan tindakan untuk agen tanpa menggunakan informasi apa pun tentang keadaan lingkungan yang disebut konteks. Bandit kontekstual memperluas multi-strategi dengan membuat keputusan keluaran bergantung pada keadaan lingkungan. Hal ini memungkinkan kami untuk mempersonalisasi setiap keputusan pada suatu situasi berdasarkan pengamatan sebelumnya. Algoritma bandit kontekstual mengamati konteks, membuat keputusan, memilih tindakan dari distribusi tindakan alternatif, dan mengamati hasil dari keputusan tersebut. Nilai fungsi penghargaan dikaitkan dengan setiap keputusan. Tujuan pembelajaran mesin adalah memaksimalkan imbalan rata-rata. Tidak seperti pembelajaran yang diawasi,

algoritma bandit kontekstual tidak memiliki semua nilai imbalan untuk setiap tindakan yang mungkin dilakukan.

Dalam pembelajaran mesin, bandit kontekstual memiliki aplikasi dalam pengoptimalan hyperparameter, pemilihan fitur, pemilihan algoritme, pembelajaran aktif, pengelompokan kolaboratif, dan pembelajaran penguatan. Bandit kontekstual manipulasi menciptakan manipulasi pada konteks dan imbalan dari bandit kontekstual. "Penyesalan" bagi pemain teori permainan membandingkan imbalan kumulatif bagi para bandit kontekstual yang bermusuhan dengan imbalan terbaik jika dipikir-pikir, yang mungkin terjadi pada kelas kebijakan. Batasan penyesalan dapat diturunkan sebagai skor keyakinan pada konsep solusi dalam pembelajaran manipulasi teoritis permainan yang diterapkan pada bandit kontekstual yang memecahkan masalah pengambilan keputusan berurutan. Seldin dkk. secara manipulasi mengkontaminasi rezim stokastik untuk bandit multi-senjata. Beberapa tuas kontrol diusulkan untuk memperbarui kecepatan pembelajaran, penyesalan empiris, dan kerugian manipulasi. Kinerja kasus terburuk dan batas penyesalan dari algoritma acak untuk bandit stokastik dalam rezim manipulasi kemudian diselidiki. House telah menghasilkan tesis tentang pendekatan teori permainan terhadap skenario multiarmed bandit. Di sini musuh adalah pengontrol yang rasional dan kompetitif yang mengurangi keuntungan pembelajar.

Analisis terhadap permainan pembelajar dan permainan balasan lawan kemudian menemukan informasi tentang eksplorasi dan eksploitasi dalam bentuk imbalan jangka panjang untuk berbagai jenis permainan. Teknik rekonstruksi matriks dan penyelesaian matriks diterapkan untuk memperkirakan keuntungan jangka panjang. Mereka memperhitungkan biaya pembelajaran yang tidak nol yang sebanding dengan biaya peluang di bidang ekonomi. Andersen dkk. menyelidiki kemampuan jaringan saraf konvolusional untuk mengekstrak fitur berguna dalam pembelajaran penguatan mendalam. Game strategi real-time dipilih sebagai domain aplikasi untuk perencanaan jangka pendek dan jangka panjang. Arsitektur Deep Q-Learning diusulkan sebagai solusinya.

Auer dkk. memperluas arsitektur Deep Q-Learning ke studi tentang strategi yang dapat menjamin hasil jangka panjang yang diharapkan dalam masalah multiarmed bandit untuk penjudi atau pemain. Tujuan seorang penjudi digambarkan sebagai memaksimalkan total hadiah melalui serangkaian uji coba di mana setiap kelompok bandit memiliki distribusi hadiah yang berbeda. Musuh mengontrol perolehan hadiah yang terkait dengan masing-masing lengan pada setiap langkah waktu. Musuh memiliki akses terhadap kekuatan komputasi tak terbatas untuk menghasilkan proses stokastik yang mendasarinya. Performa pemain diukur dalam bentuk penyesalan yaitu selisih antara hadiah kumulatif yang dicetak oleh pemain dan total hadiah yang dicetak oleh pemain terbaik.

Penyesalan seperti itu dihitung secara spesifik berdasarkan urutan imbalan yang dihasilkan oleh pihak lawan. Batas bawah dan batas atas disediakan untuk kompleksitas komputasi penyesalan algoritme dalam permainan informasi parsial. Masalah bandit yang bermusuhan seperti itu dapat dianalisis sebagai permainan matriks berulang yang tidak diketahui. Dalam permainan seperti itu, pemain tidak memiliki pengetahuan sebelumnya

tentang lawannya. Sebaliknya, musuh memainkan permainan berulang-ulang melawan pemain yang memiliki pengetahuan lengkap tentang permainan tersebut dan kekuatan komputasi yang tidak terbatas. Di sini nilai permainan bagi pemain adalah hasil terbaik yang diharapkan. Strategi acak dihitung melalui pemrograman matematika untuk mencapai hasil tersebut.

Ilyas dkk. mengintegrasikan pembuatan contoh manipulasi kotak hitam dengan optimasi bandit yang melibatkan prior pada distribusi gradien fungsi kerugian yang ditargetkan. Model ancaman blackbox seperti itu hanya dapat mengeluarkan kueri klasifikasi ke jaringan yang ditargetkan. Sebaliknya, serangan kotak putih (white-box) yang mengeksploitasi pengetahuan penuh tentang gradien fungsi kerugian yang ditargetkan untuk menciptakan contoh manipulasi. Di sini serangan yang ditargetkan dalam contoh manipulasi menyebabkan klasifikasi kelas target yang bukan kelas aslinya, sedangkan serangan yang tidak ditargetkan menyebabkan kesalahan klasifikasi secara umum. Untuk membuat serangan blackbox yang efisien dalam kueri, metode kuadrat terkecil dari pemrosesan sinyal diusulkan sebagai solusi optimal untuk masalah estimasi gradien yang menghasilkan contoh manipulasi. Memasukkan prior yang bergantung pada data dalam serangan blackbox menghasilkan solusi kueri yang efisien dibandingkan dengan yang canggih. Metode perbedaan hingga berdasarkan regresi kuadrat terkecil menghasilkan estimasi gradien informasi-teoretis dari fungsi kerugian yang ditargetkan dalam skenario serangan proyeksi penurunan gradien (PGD) yang diulang. Model kuadrat terkecil diselesaikan dengan algoritma optimasi bandit.

Untuk pemeringkatan yang dipersonalisasi dan pemodelan perhatian, Bouneffouf dkk. membuat bandit kontekstual dengan konteks terbatas terbatas pada subset fitur tetap yang dapat diakses oleh pelajar pada setiap iterasi. Fitur-fitur tersebut dirancang untuk menangani lingkungan stasioner dan non-stasioner. Masalah pembelajarannya adalah memilih subset fitur terbaik sehingga keseluruhan imbalan dimaksimalkan dengan menjelajahi ruang fitur dan ruang senjata. Dia dkk. mengusulkan pemeringkatan yang dipersonalisasi dengan tujuan yang berlawanan untuk faktorisasi matriks dalam sistem pemberi rekomendasi. Gangguan manipulasi terjadi pada penyematan vektor pengguna dan rekomendasi item dalam pemfilteran kolaboratif. Oleh karena itu, metode pelatihan baru untuk pemeringkatan yang dipersonalisasi dapat menghasilkan model pemberi rekomendasi yang kuat. Mereka dapat diperluas ke model berbasis fitur umum seperti mesin faktorisasi saraf yang mendukung berbagai skenario rekomendasi. Jadi pemeringkatan yang dipersonalisasi dengan pembelajaran adversarial dapat diterapkan dalam tugas pengambilan informasi seperti rekomendasi yang kuat, pengambilan teks, penelusuran web, menjawab pertanyaan, dan penyelesaian grafik pengetahuan.

Pembelajaran Penguatan manipulasi Game Theoretical

Penelitian pembelajaran adversarial teori permainan dapat diperluas ke pembelajaran penguatan karena fungsi tujuan teori permainan dari pembelajaran mesin adversarial dapat diinterpretasikan sebagai masalah optimasi dua tingkat yang diselesaikan dengan metode aktor-kritikus teori keputusan. Tugas klasifikasi manipulasi dengan penguatan dapat dipisahkan menjadi tugas belajar memprediksi preferensi serangan dan tugas

mengoptimalkan kebijakan operasional yang secara eksplisit mematuhi batasan operasional pada prediktor. Kemudian strategi respon terbaik musuh dihitung sebagai keputusan operasional yang diacak. Untuk mempelajari teori pembelajaran mesin yang kuat, kita dapat mengembangkan tujuan komputasi dan model inferensi statistik dalam permainan prediksi acak untuk diskriminasi, kemampuan belajar, dan keandalan algoritma manipulasi.

Kita dapat membandingkan dan membedakan pengoptimalan kotak hitam dalam pembelajaran manipulasi teoretis permainan dengan pembelajaran penguatan mendalam multi-agen untuk kemampuan generalisasi model. Di sini penelitian terhadap fungsi tujuan terbatas untuk pembelajaran manipulasi didorong oleh kemampuan dan kontrol musuh terhadap data pelatihan dan data validasi dengan mempertimbangkan skenario serangan spesifik aplikasi seperti efek pada kelas sebelumnya, pecahan sampel, dan fitur yang dimanipulasi oleh musuh. Tergantung pada tujuan, pengetahuan, dan kemampuan musuh, skenario ini juga diklasifikasikan berdasarkan pengaruh serangan, pelanggaran keamanan, dan kekhususan serangan. Masalah optimasi terbatas pada arsitektur dangkal cenderung menghasilkan algoritma komputasi yang sulit untuk estimasi kelas dan inferensi fungsi biaya adversarial. Hal ini memerlukan kebutuhan akan arsitektur pembelajaran mendalam dalam metode statistik yang memecahkan masalah optimasi dalam fungsi hasil yang merugikan. Selain kendala operasi pada kebijakan keamanan, kendala jarak dan anggaran pada fungsi biaya adversarial juga menjadi arah penelitian.

Aturan pembaruan model yang diperoleh dari skenario serangan dapat berdampak pada konvergensi proses pelatihan dalam hal trade-off antara kemampuan belajar dan ketahanan pembelajaran diskriminatif yang diusulkan. Kita dapat mengkarakterisasi masalah diskriminasi dengan adanya gangguan dalam kaitannya dengan serangkaian poin kuat di mana pengkodean data adalah jenis strategi mitigasi kesalahan khusus masalah dalam pengklasifikasi keamanan siber. Kemudian kita akan memasukkan non-linearitas dalam klasifikasi melalui representasi data dengan fungsi non-linear.

Pengaturan seperti itu juga akan memungkinkan kita untuk mengeksplorasi berbagai pilihan pengkodean variasional dari batas-batas keputusan yang dapat dipelajari. Di sini, fungsi pembayaran teoretis game mengukur pengoptimalan yang didorong oleh pemain yang meningkatkan pelatihan dan inferensi dalam pembelajaran mesin dan lingkungan yang tidak pasti. Mereka juga menjelaskan dampak lingkungan yang tidak pasti dengan mengacu pada distribusi hasil, dan, dalam arti rasionalitas teori keputusan di sekitar batasan keputusan, fungsi pembayaran memaksimalkan utilitas yang diharapkan untuk setiap pemain yang berpartisipasi dalam permainan.

Bandit kontekstual dapat dikombinasikan dengan pembelajaran manipulasi teoretis permainan untuk menganalisis kumpulan data pelatihan multimodal, diawasi dengan lemah, berisik, jarang, dan multi-terstruktur yang ditemukan dalam pembelajaran representasi pengetahuan mendalam melalui aliran dinamis dan jaringan yang kompleks. Dekomposisi bias-varians dalam fungsi imbalan yang merugikan dapat memperoleh batas penyesalan dan batas utilitas untuk jaringan pembelajaran mendalam tersebut. Selain itu, umpan balik

pengguna atau pemain dapat diintegrasikan ke dalam ukuran kinerja pembelajaran mesin sebagai metrik validasi untuk rekomendasi yang dipersonalisasi dan peringkat persaingan.

Pembelajaran manipulasi teoretis permainan dapat digunakan untuk mengeksplorasi arsitektur jaringan saraf dan fungsi biaya manipulasi dalam proses pelatihan yang menerapkan analisis data untuk tugas pemrosesan informasi siber tersebut. Di sini kerangka batas kesalahan dengan properti tanpa penyesalan untuk pembelajaran online menyediakan alat teoretis untuk menganalisis kesalahan yang dapat ditoleransi dan memperbarui aturan untuk data manipulasi yang dihasilkan. Kita juga dapat menentukan batas utilitas jaringan saraf dalam kerangka minimalisasi risiko empiris untuk pembelajaran manipulasi. Kemudian tugas pemrosesan informasi siber dapat dirumuskan sebagai masalah pembelajaran diskriminatif dengan kriteria keterpisahan yang ditandai dengan efisiensi komputasi pada data terstruktur.

Oleh karena itu, pembelajaran mendalam yang bersifat adversarial dapat menciptakan kerangka batasan kesalahan dalam aplikasi keamanan siber. Di sini permainan prediksi acak dapat merumuskan pembelajar tentang agregasi peringkat yang kuat. Kita dapat mengekspresikan ketahanan pembelajaran, keadilan, kemampuan menjelaskan, dan transparansi dengan pembelajaran adversarial teoretis permainan. Mekanisme perhatian dalam pemodelan generatif mendalam dari strategi respons terbaik musuh variasional dapat menyimulasikan dan memvalidasi lingkungan pembelajaran bagi para bandit kontekstual. Fungsi pembayaran dapat diusulkan untuk representasi pengetahuan yang dihasilkan oleh jaringan pembelajaran mendalam untuk objek dalam prediksi multimodal, multiview, dan multitask.

Dalam pembelajaran transfer dan optimalisasi stokastik atas contoh-contoh manipulasi, pembelajaran penguatan mendalam memiliki tujuan yang sama sebagai pembelajaran mendalam manipulasi teoritis permainan. Tindakan pembelajaran penguatan biasanya dinyatakan sebagai proses keputusan Markov. Ini menggunakan teknik pemrograman dinamis dalam implementasinya. Oleh karena itu, pengambilan sampel masalah dalam pembelajaran adversarial dapat berfokus pada peningkatan ketahanan pada metode rantai Markov. Pemodelan teori permainan dapat fokus pada pengintegrasian permainan Bayesian Stackelberg dan permainan Markov Stackelberg dengan pembelajaran penguatan.

Dalam pengklasifikasi keamanan siber, fungsi biaya adversarial dapat diselidiki untuk pembelajaran penguatan bahwa ketahanan terikat pada representasi adversarial. Dalam pembelajaran generatif mendalam dengan proses keputusan Markov, kita dapat membuat autoencoder variasional dinamika pengambilan sampel ulang yang digunakan dalam pembelajaran mesin adversarial sebagai alternatif metode optimasi stokastik bebas turunan dalam pemodelan teoritis permainan musuh. Optimalisasi stokastik dalam pembelajaran teori game dapat memanfaatkan permainan keamanan Markov.

Fungsi pembayaran teoretis permainan pada sistem yang kompleks dapat memperoleh manfaat dari model pembelajaran yang melibatkan sekumpulan agen otonom yang berinteraksi dalam lingkungan bersama dalam pembelajaran penguatan multi-agen.

Lingkungan multi-agen pada dasarnya tidak stasioner. Kausalitas dan stasioneritas proses keputusan Markov dapat dieksplorasi dalam pengaturan manipulasi berdasarkan prinsip-prinsip inferensi statistik seperti maksimalisasi ekspektasi, panjang deskripsi minimum, estimasi kemungkinan maksimum, dan minimalisasi risiko empiris. Mereka memiliki aplikasi dalam tugas penambangan data seperti klasifikasi, regresi, penambangan aturan asosiasi, dan pengelompokan.

Algoritme komputasi dalam teori permainan evolusi dan metode numerik dalam teori permainan diferensial dapat menambah fungsi pembayaran teoritis permainan dengan persamaan keadaan diferensial parsial dari sistem dinamis yang memodelkan interaksi kompleks dalam kontrol stokastik sebagai fungsi tujuan teoritis permainan. Kemudian prinsip pemrograman dinamis dapat digunakan untuk mempelajari sifat konvergensi permainan teoritis optima. Jaminan reliabilitas dapat dikembangkan untuk konsep solusi menurut teori kompleksitas sampel, verifikasi formal, dan automata fuzzy dalam pembelajaran adversarial. Metode variasi dan model generatif dapat mewakili manipulasi dalam konsep solusi kerugian manipulasi dan penyematan fitur dalam keamanan siber.

Kuantifikasi yang tepat atas hipotesis yang ditetapkan dalam masalah keputusan penelitian semacam itu membawa kita ke dalam berbagai masalah fungsional, masalah orakular, tugas pengambilan sampel, dan masalah optimasi dalam permainan pembelajaran manipulasi teoretis. Di sini kita dapat membandingkan solusi dengan dasar pembelajaran mesin seperti penyertaan noise dalam prosedur pengoptimalan, penyederhanaan lanskap fungsi dengan peningkatan ukuran model, skema pengoptimalan stokastik bebas turunan, dan pengambilan sampel ulang data dalam konteks algoritma pembelajaran manipulasi.

Kerangka pembelajaran manipulasi teoretis permainan dari skenario serangan berulang dan optimalisasi pertahanan kemudian akan mampu menerapkan teori permainan pada deteksi dinamika, karakterisasi, dan prediksi dalam sistem dinamis. Dinamika kompleks yang terdeteksi akan membawa kita ke prosedur pelatihan yang berlawanan untuk optimalisasi jaringan saraf dalam yang kuat. Dalam desain pengklasifikasi yang sensitif terhadap biaya, kita dapat menentukan fungsi kerugian khusus untuk menemukan tren, peringkat, perubahan, dan peristiwa dalam distribusi data yang mendasari pola dinamis yang ditambang dari data.

5.4 ALGORITMA OPTIMISASI KOMPUTASI DALAM STRATEGI PEMBELAJARAN PERMAINAN

Dalam model kuadrat terkecil yang digeneralisasi dan model linier umum untuk analitik prediktif, fungsi kerugian klasifikasi mengoptimalkan fungsi kemungkinan data yang dikondisikan kelas dari jaringan dalam yang ditargetkan. Dalam buku ini, fungsi biaya adversarial mengatur fungsi kemungkinan tersebut dengan norma, gradien, dan ekspektasi fungsi tujuan teoretis permainan yang disimpulkan pada fungsi kerugian adversarial. Jenis fungsi tujuan tersebut menentukan jenis musuh yang berpartisipasi dalam permainan prediksi dengan pengklasifikasi. Dalam buku ini kami telah mengusulkan penyelesaian musuh untuk tujuan evolusioner dan tujuan variasi dalam permainan prediksi. Nilai optimal untuk

tujuan dicari dengan algoritma evolusioner seperti algoritma genetika, algoritma simulasi anil, dan algoritma kuadrat terkecil bergantian.

Pada bagian ini kami meninjau algoritma komputasi tambahan, operator stokastik, dan kriteria konvergensi untuk optimasi komputasi dalam model pembelajaran mendalam. Studi semacam ini diharapkan dapat membawa kita pada pengacakan, konvergensi, dan paralelisasi yang lebih baik dalam penghitungan besaran langkah dan arah langkah dalam metode optimasi stokastik. Dalam merancang aturan pembaruan berulang algoritma optimasi dan penyelesaian fungsi kebugaran untuk sistem persamaan, kami tertarik pada optimasi yang kuat, optimasi numerik, dan optimasi non-linier. Selain model teoretis permainan, optimalisasi pembelajaran mendalam yang kami minati mencakup fungsi utilitas yang ditemukan dalam algoritme maksimalisasi ekspektasi, model entropi maksimum, sistem pengklasifikasi pembelajaran, mesin faktorisasi mendalam, dan model grafis probabilistik.

Fogel mengkategorikan teknik evolusi simulasi dalam optimasi stokastik jaringan saraf. Bergantung pada aspek evolusi alami (yaitu, dipandang sebagai proses optimalisasi pemecahan masalah), teknik ini disebut algoritma genetika, strategi evolusi, dan pemrograman evolusioner. Teknik-teknik ini tidak menggunakan statistik fungsi kebugaran tingkat tinggi untuk menyatu ke dalam solusi optimal. Teknik-teknik ini tidak sesensitif metode berbasis gradien terhadap gangguan manipulasi dalam fungsi kebugaran. Pirlot menjelaskan kekuatan dan kelemahan dalam simulasi anil (SA), Tabu Search (TS), dan algoritma genetika (GA). Ledesma dkk. meninjau prosedur untuk menerapkan simulasi anil secara praktis. Bandyopadhyay dkk. menggunakan simulasi anil untuk meminimalkan tingkat kesalahan klasifikasi melintasi batas keputusan dalam klasifikasi pola.

Algoritma anil deterministik diusulkan oleh Rose untuk mengoptimalkan masalah yang berkaitan dengan pengelompokan, kompresi, klasifikasi, dan regresi. Hibridisasi GA dan SA diberikan oleh Adler. SA digabungkan dengan metode pencarian lokal menjadi rantai Markov oleh Martin et al.. Kembali dkk. dan Beyer dkk. mensurvei perkembangan strategi evolusi (ES) yang memungkinkan mutasi berkorelasi pada GA. Das dkk. meninjau semua studi teoritis utama dan varian algoritma evolusi diferensial (DE) yang diterapkan pada masalah optimasi multi-tujuan, terbatas, berskala besar, dan tidak pasti. Pelikan dkk. mengusulkan pembelajaran keterkaitan antar kandidat solusi dalam komputasi evolusioner.

Zhang dkk. mensurvei masalah pembelajaran mesin dalam kerangka komputasi evolusioner. Goldberg memberikan lebih detail tentang penerapan algoritma genetika dalam pembelajaran mesin. Michalewicz membahas optimasi numerik dari operator genetika untuk mengarah pada program evolusi. Bandaru dkk. menjelaskan model deskriptif dan model prediktif untuk penambangan data dalam kumpulan data optimasi multi-tujuan. Bertsekas membahas masalah optimasi stokastik bebas turunan. Nemirovski dkk. membahas optimasi stokastik cembung-cekung dari fungsi tujuan yang diberikan dalam bentuk integral ekspektasi. Sinha dkk. meninjau solusi evolusioner untuk masalah optimasi dua tingkat. Suryan dkk. meninjau algoritma evolusi dalam teori kontrol optimal terbalik yang memiliki aplikasi dalam teori permainan.

Pengoperasian algoritma evolusioner dalam lingkungan terbatas dianalisis oleh Eiben. Cantu-Paz memberikan survei konstruksi paralel dalam algoritma genetika. Oceansek dkk. mensurvei desain untuk estimasi paralel algoritma distribusi. Sudholt memperkenalkan desain dan analisis algoritma evolusi paralel pada arsitektur CPU multicore. Algoritma genetika telah diimplementasikan dalam model pemrograman paralel yang memalukan seperti MapReduce. Whitley dkk. memberikan pedoman untuk debugging dan pengujian perhitungan evolusioner. Teorema tidak ada makan siang gratis untuk optimasi berlaku untuk perbandingan antara kriteria optimasi perhitungan evolusioner. Whitley dkk. memberikan analisis teoritis tentang masalah penelitian, fungsi tujuan, dan algoritma optimasi dalam perhitungan evolusioner. Comon dkk. mengusulkan prinsip pencarian garis yang disempurnakan (ELS) untuk menerapkan algoritma kuadrat terkecil bergantian (ALS) dalam optimasi berulang sistem persamaan non-linier yang diwakili oleh dekomposisi tensor. Analisis teoritis penyelesaian simulasi anil (SA) untuk mesin Boltzmann dan mesin Cauchy diberikan oleh Tsallis et al.

Pembelajaran Teoritis Permainan

Teori permainan algoritmik (AGT) adalah bidang penelitian yang mencakup teori permainan dan ilmu komputer. Hal ini berkaitan dengan desain dan analisis algoritma dalam lingkungan strategis. Biasanya masukan ke algoritme didistribusikan ke beberapa pemain atau agen yang mempunyai kepentingan dalam keluaran algoritme. Aspek analisis AGT menerapkan alat teori permainan seperti dinamika respon terbaik dalam implementasi dan analisis algoritma. Aspek desain AGT adalah tentang pemodelan komputasi sifat teoritis permainan dan pola algoritmik dalam desain dan peningkatan kompleksitas algoritma. Interaksi berbasis internet antara agen komputasi dapat dimodelkan dengan keseimbangan teoritis permainan yang terkait dengan pemodelan analisis data. Pilihan sosial komputasional adalah bidang penelitian yang memperluas model teoritis permainan ke sistem multi-agen yang menggabungkan preferensi individu agen dalam mekanisme online.

Strategi Pengacakan dalam Pembelajaran Game Theoretical Adversarial

Grunwald dkk. menunjukkan kesetaraan antara pemodelan entropi maksimum dan meminimalkan kemungkinan kerugian terburuk dari teori keseimbangan permainan zero-sum untuk fungsi kerugian dan masalah keputusan. Entropi relatif umum dengan kondisi keteraturan diusulkan untuk menganalisis pengklasifikasi kuat yang meminimalkan perbedaan antar distribusi. Teorema minmax kemudian diusulkan untuk divergensi Kullback-Leibler antara data pelatihan dan distribusi data manipulasi yang diperlakukan sebagai keluarga distribusi eksponensial umum. Hal ini memberikan interpretasi teoretis keputusan tentang prinsip entropi maksimum di mana fungsi kerugian adversarial tidak hanya dianggap sebagai skor logaritmik. Definisi teoretis keputusan tentang perbedaan atau entropi relatif antara distribusi probabilitas untuk pelatihan dan data manipulasi menggeneralisasi divergensi Bregman terhadap fungsi kerugian dalam pembelajaran mesin. Pemodelan entropi maksimum dianggap sebagai versi pengklasifikasi Bayes yang kuat.

Pengklasifikasi yang sensitif terhadap biaya dalam pembelajaran mesin telah memperoleh manfaat dari properti permainan zero-sum dalam teori permainan. Algoritme

pembelajaran manipulasi telah melakukan perbaikan pada formulasi permainan minmax untuk menghasilkan pengklasifikasi yang kuat. Rezek dkk. menunjukkan kesetaraan antara kesimpulan yang diambil pada observasi sebelumnya yang dibuat dalam teori permainan dan pembelajaran mesin. Respons halus terbaik dalam permainan fiktif yang diulang-ulang dikontraskan dengan metode inferensi Bayesian dalam pembelajaran mesin yang terintegrasi pada distribusi manipulasi, bukan rata-rata empiris.

Kemudian teori permainan digunakan dalam analisis dan desain algoritma pembelajaran variasional. Untuk mengelompokkan campuran distribusi, algoritme pembelajaran variasional menunjukkan sifat konvergensi yang kuat dan memperbarui aturan. Solusi yang diusulkan berkaitan erat dengan perkembangan model grafis probabilistik. Jadi model grafis probabilistik dapat menghasilkan algoritma yang efisien untuk menghitung keseimbangan Nash dalam game multipemain besar untuk pembelajaran mesin yang diawasi. Secara umum, desain algoritme pembelajaran mesin untuk lingkungan stasioner dalam lingkungan akademis yang ideal dapat memperoleh manfaat dari analisis teoretis permainan skenario non-stasioner yang sering ditemukan dalam aplikasi dinamis teknik pembelajaran mesin di dunia nyata.

Pembelajaran Mendalam Mesin Adversarial dalam Game yang Kuat

Bowling dkk. meneliti pembelajaran penguatan multi-agen menggunakan kerangka permainan stokastik. Permainan stokastik diperlakukan sebagai perpanjangan dari proses pengambilan keputusan Markov ke banyak agen. Pembelajaran penguatan kebijakan agen di hadapan agen pembelajaran lainnya dianalisis untuk mengetahui sifat pembelajaran yang disebut rasionalitas dan konvergensi. Untuk berperan dalam mencapai solusi keseimbangan, sifat rasionalitas mengharuskan pemain untuk mengadopsi strategi respons terbaik untuk mempelajari kebijakan mengingat pemain yang tersisa telah memainkan strategi stasioner.

Properti konvergensi memastikan bahwa semua pemain yang berpartisipasi dalam permainan stokastik pada akhirnya berada dalam kebijakan stasioner yang dikondisikan pada algoritma pembelajaran pemain lain. Jika kedua sifat ini terpenuhi, maka semua pemain dijamin akan mencapai keseimbangan Nash. Setiap keadaan dalam permainan stokastik dipandang sebagai permainan matriks. Para pemain teori permainan bertransisi dari satu permainan matriks ke permainan matriks lainnya setelah menerima hadiah yang ditentukan oleh tindakan bersama mereka. Algoritme pembelajaran penguatan yang dipertimbangkan adalah pembelajar agen tunggal, pembelajar tindakan bersama, dan d minimax-Q. Kecepatan pemelajaran variabel digunakan untuk memperbarui estimasi nilai Q dalam kebijakan pendakian bukit. Hal ini sebanding dengan algoritma pembobotan acak dalam teori permainan evolusi yang mendistribusikan ulang bobot di antara para ahli yang salah.

Permainan stokastik dapat didefinisikan sebagai kumpulan permainan bentuk normal yang dimainkan berulang kali oleh agen. Hal ini dapat direpresentasikan sebagai robot probabilistik di mana negara bagian adalah permainannya dan label transisi adalah pasangan tindakan-hasil bersama. Permainan berulang seperti dilema tahanan yang berulang adalah permainan stokastik dengan hanya satu keadaan. Proses pengambilan keputusan Markov adalah permainan stokastik dengan hanya satu pemain. Strategi deterministik menentukan

pilihan tindakan untuk pemain teori permainan. Strategi campuran adalah distribusi probabilitas atas strategi deterministik. Ekuilibrium Nash dari permainan sekuensial telah diperluas ke konsep solusi untuk permainan stokastik yang disebut keseimbangan sempurna Markov.

Permainan stokastik dapat digabungkan dengan permainan Bayesian untuk mencapai keseimbangan Bayesian Nash. Permainan stokastik dua pemain pada grafik terarah digunakan untuk memodelkan sistem diskrit yang beroperasi di lingkungan manipulasi yang tidak diketahui. Konfigurasi sistem diskrit dan lingkungan manipulasi nya direpresentasikan sebagai simpul dari grafik berarah. Transisi antar node berhubungan dengan tindakan bersama dalam sistem diskrit. Jalur dalam grafik berarah berhubungan dengan eksekusi sistem operasional. Berbagai konsep solusi seperti kesetimbangan posisi, kesetimbangan stasioner, kesetimbangan acak, dan kesetimbangan keadaan terbatas dimungkinkan dalam permainan stokastik.

Lippi dkk. mengusulkan model dan algoritma dari pembelajaran relasional statistik (SRL) sebagai alat untuk analisis dan desain pemodelan teoritis permainan dalam permainan stokastik. Jadi logika orde pertama dan model grafis probabilistik seperti jaringan Bayesian atau jaringan Markov dapat mewakili ketidakpastian dalam permainan karena ketergantungan antar variabel acak. Algoritme inferensi statistik dari SRL seperti metode variasional dapat menemukan keseimbangan Nash dan solusi optimal Pareto untuk masalah pembelajaran adversarial teoretis permainan. Di sini keseimbangan Pareto-optimal untuk permainan adalah serangkaian profil strategi untuk para pemain di mana tidak ada yang dapat meningkatkan hasil mereka tanpa mengurangi hasil pemain lain.

Sebaliknya, keseimbangan Nash tercapai ketika profil yang dipilih oleh setiap pemain dalam permainan merupakan respons terbaik terhadap profil yang dipilih oleh pemain yang tersisa. Ekuilibrium Pareto-optimal dan Nash dapat diperluas ke game multipemain. Algoritma pembelajaran struktur dari SRL seperti jaringan logika Markov dapat menghasilkan klausa logika probabilistik yang dapat ditafsirkan untuk menggambarkan strategi musuh pada tingkat tinggi hingga manusia. Di sini permainan grafis menerapkan model teori permainan ke grafik kombinatorial dalam pembelajaran mesin dan pencarian pohon Monte Carlo dapat memprediksi evolusi manipulasi teoritis permainan. Game grafis menganggap pemain sebagai node dalam grafik dan edge mewakili interaksi mereka. Jadi, imbalan seorang pemain bergantung pada tetangganya, bukan seluruh pemain dalam permainan.

Hal ini menghasilkan beberapa matriks pembayaran lokal untuk seorang pemain. Formalisme logika dalam SRL seperti pemrograman logika induktif dapat mengatasi pembelajaran representasi pengetahuan pada permainan untuk menggambarkan domain yang diminati dalam teori permainan seperti strategi, aliansi, aturan, hubungan, dan ketergantungan antar pemain. Mereka juga dapat menemukan informasi tentang lingkungan eksternal untuk pembelajaran manipulasi. Penalaran probabilistik juga berguna untuk menangani informasi yang hilang atau tidak lengkap untuk pengambilan keputusan dalam pemodelan teoretis game untuk pembelajaran mesin. Jadi SRL dapat diterapkan dalam skenario pengambilan keputusan melalui pembelajaran penguatan dan pemodelan musuh.

Logika Markov dalam pembelajaran mesin adversarial teoretis permainan memungkinkan kita memodelkan pengetahuan adversarial dalam bentuk predikat logika tentang bukti (fakta yang diketahui) atau kueri (fakta yang akan disimpulkan). Hal ini dapat diperluas ke kerangka teori keputusan yang melampirkan fungsi utilitas ke klausa orde pertama dalam jaringan keputusan logika Markov. Algoritma pemaksimalan ekspektasi dapat dibangun untuk menyimpulkan nilai predikat logika. Sifat relasionalnya dapat dimanfaatkan untuk memodelkan algoritme klasifikasi kolektif dalam game multipemain dengan strategi yang dapat ditafsirkan untuk aplikasi dunia nyata.

Aghassi dkk. mengusulkan optimasi kuat bebas distribusi untuk mengatasi ketidakpastian hasil dalam permainan informasi yang tidak lengkap. Kesetimbangan optimasi yang kuat dianalisis untuk permainan terbatas dengan kumpulan ketidakpastian hasil polihedral yang terbatas. Keseimbangan seperti itu dapat dikontraskan dengan permainan terbatas non-kooperatif, gerakan simultan, satu pukulan, dengan informasi lengkap yang mengarah ke keseimbangan Nash. Pada ekuilibrium Nash, para pemain teoritis permainan memaksimalkan hasil yang diharapkan sehubungan dengan distribusi probabilitas yang diberikan oleh ruang strategi campuran.

Model utilitas yang diharapkan dalam kasus terburuk sangat cocok untuk menganalisis situasi teori keputusan yang ditandai dengan pemodelan ketidakpastian seputar penilaian risiko yang merugikan dalam informasi distribusi yang tersedia untuk pembelajaran mesin sebagai kumpulan data pelatihan, pengujian, dan validasi. Di sini sumber ketidakpastian dalam pemodelan disebabkan oleh ketidakpastian pembayaran masing-masing pemain berdasarkan rangkaian tindakan, ketidakpastian dalam perilaku pemain, dan distribusi probabilitas sebelumnya di sekitar konfigurasi beberapa pemain. Untuk menyelesaikan permainan informasi yang tidak lengkap, digunakan kriteria keputusan bebas distribusi penyesalan minimax untuk optimalisasi pembelajaran online.

Permainan tangguh yang diusulkan oleh Aghassi dkk. sebanding dengan game online semacam itu. Kriteria validasi kinerja untuk pembelajaran mesin yang dirancang dengan pemodelan teoritis permainan tersebut kemudian dinyatakan dalam bentuk hasil terburuk yang diharapkan pemain. Kesetimbangan teoretis permainan dirumuskan sebagai proyeksi solusi komponen pereduksi dimensi untuk sistem persamaan dan pertidaksamaan multilinear. Mereka dapat diperluas ke permainan terbatas yang kuat yang memiliki informasi pribadi tambahan tentang keyakinan pemain. Cohen dkk. melakukan analisis nilai Shapley multiperturbasi untuk memperkirakan kegunaan fitur dalam seleksi maju dan eliminasi mundur dari algoritma pemilihan fitur. Fitur yang difilter mampu mengoptimalkan ukuran kinerja pada data yang tidak terlihat seperti akurasi, tingkat kesalahan, dan area di bawah kurva ROC.

Dengan demikian pemodelan teoritis permainan pemilihan fitur dapat melakukan reduksi dimensi untuk membuat subset fitur guna meningkatkan kinerja prediksi, mengurangi kebutuhan pengukuran dan penyimpanan, mengurangi waktu pelatihan dan prediksi, memberikan pemahaman yang lebih baik tentang distribusi data yang mendasarinya, dan menghasilkan visualisasi data yang relevan. fitur. Nilai Shapley suatu fitur mengukur

kinerjanya dalam subset fitur. Teori himpunan terbatas dikombinasikan dengan estimasi nilai Shapley untuk menemukan algoritma yang efisien untuk mengekstraksi fitur kuat dalam masalah klasifikasi. Memangkas fitur yang tidak relevan mengurangi kesalahan generalisasi pengklasifikasi. Distribusi kontribusi fitur digunakan untuk memandu algoritma pemilihan fitur. Kebisingan yang merugikan dapat memainkan peran utama dalam memanipulasi distribusi tersebut untuk menghasilkan pengklasifikasi yang berperforma rendah.

Matahari dkk. mengusulkan kerangka kerja berbasis teori permainan kooperatif untuk mengevaluasi kekuatan diskriminatif setiap fitur dalam konteks fitur yang saling terkait untuk filter pemilihan fitur. Filter seperti itu dapat diintegrasikan dengan algoritma pembelajaran apa pun untuk menghasilkan pengklasifikasi yang efisien. Solusi untuk permainan kooperatif membangun nilai bagi setiap pemain untuk menciptakan fungsi karakteristik yang mengukur kontribusi pemain terhadap permainan. Informasi timbal balik bersyarat adalah mekanisme pembobotan ulang untuk mengevaluasi relevansi fitur yang bergantung pada kelas terhadap subset fitur yang telah dipilih sebelumnya dalam algoritma pembelajaran.

Indeks kekuatan Banzhaf ditetapkan untuk setiap fitur berdasarkan kontribusi marjinalnya terhadap sifat korelasi intrinsik antar fitur seperti kausalitas, saling ketergantungan, dan independensi. Subset fitur tersebut kemudian digunakan untuk menghitung proporsi koalisi pemenang dalam penghitungan fungsi pembayaran. Pemrograman dinamis digunakan untuk mengimplementasikan permainan kooperatif dengan kompleksitas waktu yang berkurang. Chalkiadakis dkk. memberikan survei algoritma komputasi dalam teori permainan kooperatif. Penekanan khusus ditempatkan pada konsep solusi komputasi yang efisien dan representasi kompak untuk permainan.

Tinjauan tentang algoritma pemaksimalan kesejahteraan untuk pembelajaran manipulasi teoritis permainan yang membentuk koalisi dan tawar-menawar juga disediakan. Pengoptimalan kombinatorial dalam masalah pembelajaran mesin dapat memperoleh manfaat dari diskusi tentang permainan subgraf terinduksi, permainan pohon rentang biaya minimum, dan permainan aliran jaringan. Pendekatan pemrograman dinamis untuk menghasilkan struktur koalisi yang optimal juga dibahas. Mereka dapat disesuaikan dengan algoritma kapan saja yang menghasilkan solusi yang lebih baik secara bertahap dengan lebih banyak waktu atau sumber daya komputasi.

Garg dkk. mengembangkan pemodelan teoritis permainan untuk pengelompokan fitur. Fitur dipandang sebagai pemain rasional dalam permainan koalisi dan koalisi ditafsirkan sebagai kelompok. Partisi stabil Nash (NSP) adalah konsep solusi dari teori permainan koalisi yang digunakan untuk menyediakan konfigurasi pengelompokan akhir dari fitur-fitur tersebut. Properti yang diinginkan dalam cluster dapat dipilih dengan mengacu pada berbagai fungsi pembayaran teoritis permainan. NSP ditemukan dengan menyelesaikan program linier bilangan bulat (ILP). Pendekatan pengelompokan hierarki kemudian diusulkan untuk menskalakan pengelompokan dengan partisi grafik. Semua fitur yang dipilih dalam sebuah cluster relevan dan saling melengkapi satu sama lain. Untuk melakukan ekstraksi fitur menggunakan teknik clustering, diusulkan juga pemeringkatan fitur dari cluster fitur.

Bektor dkk. meninjau pemrograman matematika fuzzy dalam pemodelan teoritis permainan. Himpunan fuzzy dapat diterapkan pada bidang penelitian seperti pemrograman matematika dan teori permainan matriks yang terjadi pada antarmuka teori permainan dan teori keputusan. Lingkungan fuzzy dapat memberikan generalisasi pada tujuan teoritis permainan penyelesaian pemrograman linier dan kuadrat dalam permainan matriks terbatas dalam permainan dua pemain bukan jumlah nol yang memiliki tujuan fuzzy. Permainan matriks dengan hasil fuzzy dapat memodelkan masalah pemrograman linier multi-objektif dalam pembelajaran adversarial. Beberapa konsep solusi untuk permainan matriks fuzzy tersebut kemudian dijelaskan.

Ketidakjelasan fungsi keputusan untuk pengklasifikasi adversarial dapat dimodelkan sehubungan dengan tujuan pembelajaran, lingkungan, dan batasan adversarial. Hal ini mengarah pada masalah pemrograman matematika fuzzy dalam pembelajaran adversarial teoritis permainan. Misalnya, hubungan preferensi fuzzy dapat digunakan untuk algoritma pembelajaran representasi pengetahuan melalui kumpulan data multimodal untuk akhirnya memecahkan masalah pemrograman matematika dengan kendala modalitas yang merumuskan model teoritis permainan dalam pembelajaran mesin. Algoritme komputasi harus dikembangkan untuk menemukan solusi optimal untuk masalah optimasi fuzzy dalam pembelajaran mesin adversarial teoritis permainan.

Perc dkk. melakukan survei kerjasama dalam teori permainan evolusi untuk memecahkan masalah yang disebut dilema sosial yang mewakili stokastisitas interaksi antara para pemain teori permainan. Evolusi strategi, pendukung kerja sama, dan aturan ko-evolusioner digunakan untuk mengekspresikan munculnya kerja sama dan pembelotan dalam permainan evolusi. Interaksi dinamis antar pemain dapat dipelajari dengan aturan ko-evolusioner pada jaringan kompleks yang mewakili jaringan interaksi, pertumbuhan populasi data, mobilitas pemain, dan penuaan pemain. Ficici dkk. memperkenalkan pemodelan teori permainan dalam mekanisme memori ko-evolusi. Kumpulan ciri-ciri ingatan yang menonjol direpresentasikan sebagai strategi campuran.

Memori tersebut mewujudkan solusi proses ko-evolusi yang diperoleh pada keseimbangan Nash. Memori dapat mengalami keterbatasan sumber daya selama pelatihan. Memori dan penyimpangan dalam ko-evolusi ditafsirkan sebagai kesalahan pengambilan sampel dan bias variasional dalam pemodelan teoretis permainan. Sensitivitas dan kontinjensi seputar evaluasi fungsi kebugaran proses ko-evolusi dapat menyebabkan sistem pembelajaran mesin mempelajari, melupakan, dan mempelajari kembali ciri-ciri memori secara siklik. Konsep solusi dalam teori permainan kemudian mewakili kumpulan ciri-ciri memori yang termasuk dalam himpunan yang diinginkan atau benar. Mekanisme “Memori Nash” yang diusulkan mengumpulkan kumpulan sifat-sifat sebagai strategi respons terbaik.

Strategi ekuilibrium Nash memberikan solusi respons terbaik yang menyatakan tingkat keamanan permainan evolusi sebagai ekspektasi hasil tertinggi yang dicapai oleh semua pemain yang bertindak secara kolektif. Tingkat keamanan ini disebut juga dengan nilai permainan. Heuristik pencarian dirancang pada populasi yang berevolusi bersama yang mampu meringankan beban populasi dalam merepresentasikan solusi dan berkonsentrasi

pada pencarian untuk meningkatkan solusi yang diwakili oleh memori. Algoritma waktu polinomial digunakan untuk menyelesaikan permainan zero-sum dengan pemrograman linier. Ruang strategi untuk menemukan ingatan Nash mungkin terbatas atau tidak terbatas, dapat dihitung, atau tidak dapat dihitung.

Herbert dkk. mengusulkan pemodelan teori permainan untuk pembelajaran kompetitif dalam peta pengorganisasian mandiri (SOM). Fokus proses pelatihan pada clustering berbasis SOM adalah mencari neuron yang paling mirip dengan vektor masukan. Ekstensi GTSOM yang diusulkan mengevaluasi kualitas SOM secara keseluruhan dengan mencapai posisi optimal secara global menggunakan teori permainan untuk mengusulkan aturan pembaruan dinamis dan adaptif pada bobot neuron yang mampu memperhitungkan ketidaksesuaian kepadatan dalam masalah pengelompokan. Cluster dideskripsikan dalam bentuk data masukan aktual dan neuron yang terkait dengan data tersebut. Teori permainan mampu mengurutkan neuron untuk menentukan neuron yang memberikan peningkatan kualitas SOM terbesar berdasarkan jarak dari vektor masukan.

Pengukuran kualitas tambahan pada neuron juga dapat diterapkan untuk mempertimbangkan peta fitur terkait yang diekstraksi dari data. Strategi teoretis permainan diusulkan untuk menyesuaikan kecepatan pembelajaran SOM sedemikian rupa sehingga vektor masukan akan memiliki kemungkinan lebih besar untuk lebih dekat ke neuron berbeda pada iterasi berikutnya dari algoritma pelatihan. Serangkaian tindakan teoritis permainan merinci lingkungan pengelompokan dan kepadatan untuk membedakan atau mengurangi kelompok yang diinginkan. Pelatihan berakhir jika SOM telah mencapai ambang batas yang ditentukan pengguna untuk preferensi kualitas pengelompokan.

Schuurmans dkk. menyelidiki hubungan antara metode pembelajaran mendalam yang diawasi dan teori permainan. Strategi tanpa penyesalan dalam pemodelan teori permainan ditemukan sebagai metode pelatihan stokastik yang efektif untuk masalah pembelajaran yang diawasi. Pencocokan penyesalan diusulkan sebagai alternatif penurunan gradien untuk mengoptimalkan kinerja stokastik pembelajaran mendalam yang diawasi secara efisien. Proses pembelajaran yang diawasi melalui jaringan saraf asiklik terarah dengan fungsi aktivasi cembung yang dapat dibedakan dinyatakan sebagai permainan gerak simultan dengan tindakan dan utilitas pemain sederhana.

Pemain memilih tindakan mereka secara independen dari tindakan yang diambil oleh pemain lain. Penyesalan kumulatif untuk setiap pemain ditentukan berdasarkan fungsi utilitas yang diharapkan. Pakar domain dan alam juga dapat dipertanggungjawabkan dalam pemetaan strategi dan tindakan pembelajar. Korespondensi yang erat ditemukan antara pembelajaran online cembung dan permainan zero-sum dua orang. Algoritma bobot eksponensial dan pencocokan penyesalan diusulkan sebagai algoritma pelatihan terbatas untuk pembelajaran yang diawasi. Hasil pelatihan tentang batas penyesalan, kriteria konvergensi, dan optima global dari algoritma pelatihan yang dibatasi dibandingkan dengan proyeksi penurunan gradien stokastik dan penurunan gradien stokastik.

Algoritme pelatihan yang dibatasi ternyata sangat kompetitif dalam ruang fitur renggang berdimensi tinggi dalam jaringan pembelajaran yang diawasi. Ekuilibrium Nash

dijamin menjadi salah satu optimal lokal jika bukan optimal global untuk pelatihan jaringan saraf dalam. Algoritme pencocokan penyesalan dalam evaluasi ditemukan menghasilkan kesalahan kesalahan klasifikasi yang lebih rendah dibandingkan metode pembelajaran mendalam standar. Namun, teori yang diajukan tidak berlaku untuk jaringan saraf dengan fungsi aktivasi yang tidak lancar dalam beberapa lapisan tersembunyi. Untuk mencapai global optima, Oliehoek dkk. memodelkan jaringan manipulasi generatif (GAN) menurut permainan terbatas dalam strategi campuran. Konsep solusi yang diusulkan secara monoton menyatu ke keseimbangan Nash yang dibatasi sumber daya yang merupakan titik pelana dalam strategi campuran.

Konsep solusi dapat dibuat lebih akurat dengan sumber daya komputasi tambahan untuk memperkirakan perhitungan respons terbaik. Algoritme pelatihan yang diusulkan untuk pemodelan generatif mendalam mampu menghindari masalah umum seperti keruntuhan mode, degenerasi mode, penghilangan mode, dan lupa mode. Metode teori permainan yang diusulkan disebut Parallel Nash Memory. Hal ini dapat dieksploitasi untuk menghasilkan peningkatan dalam metrik pelatihan yang kuat dari kinerja jaringan pengklasifikasi/generator. Diskriminator dan model generator kemudian dapat diperbarui sesuai dengan strategi respons terbaik pada setiap iterasi.

Mereka dapat secara eksplisit membatasi strategi yang diperbolehkan dengan mesin keadaan terbatas. Model tangguh yang dihasilkan menghasilkan kinerja generatif yang lebih baik dengan kompleksitas total yang sama dan lebih mendekati keseimbangan Nash global. Mereka dapat diperluas dengan permainan polimatriks zero-sum dan permainan tereduksi dengan data manipulasi yang memandu pelatihan. Hsieh dkk. juga mengusulkan strategi pelatihan bagi GAN untuk menemukan keseimbangan campuran Nash. Metode pengambilan sampel lebih lanjut diusulkan untuk menyelesaikan permainan strategi campuran. Skema pengambilan sampel perkiraan rata-rata yang diusulkan dapat menambah kerangka optimalisasi global untuk pembelajaran manipulasi teoritis permainan.

Secara khusus, skema pengambilan sampel perkiraan rata-rata untuk permainan bi-affine diselidiki untuk menyediakan algoritma pelatihan praktis untuk GAN. Pelatihan yang kuat ini memformulasi ulang distribusi GAN pada strategi terbatas sebagai ukuran probabilitas pada kumpulan parameter berkelanjutan. Metode pengambilan sampel yang disebut entropic mirror descending memperkirakan ukuran probabilitas tersebut dengan cara yang mudah diatur. Oleh karena itu, pelatihan yang kuat memformulasi ulang dinamika pelatihan algoritma berbasis gradien ke dalam program minmax yang diselesaikan dengan pemrograman matematika dan teori permainan algoritmik. Dalam percobaan, optima stasioner yang ditemukan oleh algoritma berbasis gradien seperti SGD, Adam, dan RMSProp ditemukan bukan minmax optima secara lokal atau global. Hal ini mengarah pada pengembangan lebih lanjut dalam intuisi pengoptimalan non-cembung yang diterapkan pada validasi pembelajaran mesin.

Tembine dkk. menyajikan interaksi antara permainan yang kuat secara distribusi dan jaringan manipulasi generatif yang mendalam. Perbedaan Bregman antara distribusi data adversarial dan pelatihan dibuat untuk menghindari penggunaan turunan kedua dari fungsi

tujuan dalam algoritma optimasi yang diterapkan pada GAN. GAN diformulasikan sebagai game yang kuat secara distribusi dalam lingkungan multi-agen yang bermusuhan. Para pemain dalam permainan bentuk strategis adalah unit neuron. Drama tersebut adalah beban yang dipelajari. Fungsi tujuan teoritis permainan merupakan fungsi kerugian yang diperoleh dari ketidaksesuaian antara keluaran dan pengukuran data nyata. Tingkat konvergensi dari algoritma pembelajaran mendalam yang diusulkan diperoleh dengan menggunakan perkiraan rata-rata.

Pembelajaran mean-field dipandang sebagai kelas kandidat algoritme yang akan diselidiki untuk pembelajaran mendalam berdimensi tinggi dengan musuh teoretis permainan yang bertindak sebagai pengambil keputusan. f-divergence dan metrik Wasserstein digunakan dalam evaluasi eksperimental untuk menemukan ketidaksesuaian antara data yang dihasilkan dan data sebenarnya. Jadi lapisan tersembunyi dalam jaringan saraf dipandang sebagai lingkungan interaktif dinamis yang direpresentasikan sebagai permainan. Fungsi keluaran multimodal dalam jaringan saraf dalam menimbulkan kesulitan lebih lanjut dalam permainan optimalisasi teoritis parameter strategis dan saling bergantung dalam algoritma pelatihan jaringan saraf.

Ketidakselarasan antara pembaruan komponen pelatihan jaringan saraf memotivasi perlunya imbalan teori permainan dalam algoritme pelatihan seperti propagasi balik kesalahan, penurunan gradien stokastik, dan algoritme berbasis mean-field atau populasi (seperti genetika, gerombolan, dan simulasi anil). Estimasi nilai gradien yang diharapkan dalam metode berbasis turunan lebih lanjut memerlukan metode pengambilan sampel seperti pengambilan sampel Monte Carlo, pembelajaran penguatan, atau pengambilan sampel berbasis populasi yang diintegrasikan dengan permainan manipulasi multi-agen aksi berkelanjutan.

Dimulai dari berbagai dimensi sistem dinamis, seperti pembelajaran mendalam yang strategis atau pembelajaran teori permainan yang mendalam juga berguna dalam mengatasi hiper-parametrisasi, kutukan dimensi, dan propagasi kesalahan. Konsep solusi teori permainan yang relevan untuk pembelajaran mendalam strategis berbasis model berbasis data di berbagai model ancaman mencakup keseimbangan Nash, solusi Stackelberg, Pareto optima, solusi Berge, solusi tawar-menawar, dan keseimbangan berkorelasi.

Optimasi Kuat dalam Pembelajaran Tantangan

Xu dkk. mengatur mesin vektor dukungan dalam formulasi optimasi probabilistik yang kuat untuk klasifikasi. Pengoptimalan yang kuat meminimalkan kemungkinan kesalahan empiris terburuk pada distribusi dasar sebenarnya dari sampel data pelatihan yang dicampur dengan non-i.i.d. gangguan (berpotensi merugikan). Pengoptimalan tangguh yang diusulkan menawarkan perlindungan terhadap kebisingan, membantu mengontrol overfitting, dan menghasilkan kinerja generalisasi. Istilah regularisasi menyelesaikan himpunan ketidakpastian tipe non-kotak. Pengklasifikasi dengan peluang terbatas adalah hasil dari optimasi yang kuat. Ini adalah pengklasifikasi dengan batasan probabilistik pada tingkat kesalahan klasifikasi.

Pengaturan Bayesian memilih koefisien regularisasi tanpa validasi silang. Penelitian ini kontras dengan regularisasi pengklasifikasi yang membatasi kompleksitas kelas fungsi dalam pendekatan minimalisasi risiko struktural PAC terhadap klasifikasi. Kekokohan disebabkan oleh optimasi minmax yang dilakukan pada semua kemungkinan gangguan. Batasan ketahanan pada sampel yang rusak mampu menangani non-i.i.d. data di mana sampel pelatihan dan sampel pengujian diambil dari distribusi yang berbeda, atau beberapa musuh memanipulasi sampel untuk mencegahnya diberi label dengan benar. Hasilnya dibandingkan dengan statistik yang kuat seperti pendekatan fungsi pengaruh untuk penduga regresi atau algoritma klasifikasi yang dibangun berdasarkan gangguan kecil pada model statistik yang terdiri dari fungsi kerugian yang tidak mulus.

Dalam desain algoritma, optimasi kuat yang diusulkan memiliki keuntungan tambahan yaitu memperkuat algoritma pembelajaran ketika sifat gangguan diketahui secara apriori atau dapat diperkirakan dengan baik. Bukti konsistensi statistik pada pembelajaran yang kuat dalam ruang sampel menggantikan entropi metrik, dimensi VC, dan kondisi stabilitas dalam ruang fitur untuk mesin vektor pendukung dengan kondisi ketahanan pada kesalahan klasifikasi yang diharapkan dan kerugian yang diatur. Pandangan pengklasifikasi yang kuat seperti itu mampu memperoleh batas kompleksitas sampel untuk kelas algoritma yang luas dalam pembelajaran yang diawasi. Ini membentuk hubungan eksplisit antara regularisasi dan ketahanan dalam klasifikasi pola.

Dalam ruang fitur biner, Li et al. mewakili musuh yang rasional dan berorientasi pada tujuan berdasarkan pemrograman linier bilangan bulat campuran dengan pembangkitan kendala. Kerangka kerja pelatihan ulang yang berulang kemudian diusulkan untuk meminimalkan kerugian yang merugikan dalam serangan penghindaran. Model teoritis permainan yang dihasilkan disebut model permainan multi-musuh Stackelberg. Fungsi biaya adversarial baru diusulkan untuk memungkinkan substitusi silang fitur yang membuat trade-off antara pemilihan fitur melalui regularisasi yang jarang dan penghindaran adversarial.

Masalah optimasi bi-level yang diusulkan tidak memerlukan modifikasi dalam algoritma pembelajaran. Fungsi risiko manipulasi pada data pelatihan memodelkan kumpulan musuh. Ini menghitung risiko empiris pembelajaran mesin dalam situasi manipulasi. Keseimbangan Stackelberg yang bukan jumlah nol ditemukan antara pembela tunggal (pengklasifikasi) dan banyak pengikut (penghindar) yang beroperasi dalam batasan penganggaran berbiaya tinggi. Regularisasi kekokohan algoritma pembelajaran sebanding dengan statistik kuat yang dihitung pada kontaminasi data dalam kasus terburuk untuk fitur diskrit dan berkelanjutan.

Vorobeychik dkk. mencirikan skema pengacakan optimal dalam klasifikasi manipulasi. Pengklasifikasi bertindak sebagai pembela terhadap rekayasa balik dan manipulasi pengklasifikasi yang merugikan. Dalam eksperimen tersebut, kebijakan optimal pembela HAM adalah mengacak secara seragam untuk serangan yang ditargetkan setelah mengabaikan keakuratan klasifikasi dasar, atau tidak mengacak sama sekali tetapi memilih pengklasifikasi yang lebih baik untuk kebenaran klasifikasi yang diamati (atau disimpulkan) dan pertahanan terhadap serangan sembarangan. Mekanisme pertahanan seperti ini menarik dalam teknik

pembelajaran mesin untuk keamanan siber (atau fisik) seperti deteksi intrusi, pemfilteran spam, dan pembuatan malware yang dibingkai sebagai tugas prediksi. Dalam domain keamanan seperti itu, musuh akan secara aktif melemahkan pengklasifikasi dengan penghindaran dan sabotase yang mengarah pada kesalahan pola dan label klasifikasi.

Pengklasifikasi manipulasi kemudian dapat memanfaatkan pemodelan teoritis permainan dari interaksi pembelajar-penyerang, mempelajari kompleksitas algoritmik dari skenario serangan manipulasi, dan mengusulkan pengklasifikasi berbasis pengacakan. Lebih lanjut, rekayasa balik berbasis kueri manipulasi untuk mempelajari pengklasifikasi linier secara efisien yang digunakan oleh pembela untuk memecahkan masalah optimasi cembung di kelas yang dapat dipelajari juga diusulkan. Ini memiliki penerapan dalam domain keamanan siber berbasis pengacakan seperti target bergerak atau skema klasifikasi pertahanan dinamis. Di sini evaluasi eksperimental menemukan bahwa semakin baik kinerja dasar pengklasifikasi, semakin buruk kinerjanya setelah serangan yang ditargetkan seperti spear phishing. Tingkat kesalahan setelah serangan yang ditargetkan mengeksploitasi kesalahan klasifikasi tidak secara langsung bergantung pada tingkat kesalahan dasar pengklasifikasi yang menentukan postur keamanan operasional.

Hashimoto dkk. mengusulkan optimalisasi risiko yang kuat secara distribusi. Hal ini meminimalkan risiko terburuk pada semua distribusi yang mendekati distribusi empiris yang mewakili data pelatihan. Strategi mitigasi risiko mengatasi pengaruh kelas minoritas dalam perhitungan kerugian dan akurasi rata-rata yang disebut disparitas representasi. Hal ini diamati dalam pengenalan wajah, identifikasi bahasa, penguraian ketergantungan, penandaan part-of-speech, sistem rekomendasi, teks video, pengenalan ucapan, dan terjemahan mesin. Pengklasifikasi mitigasi risiko cenderung mencapai keadilan dalam pembelajaran yang diawasi atas label yang dilindungi melalui kalibrasi batasan ke dalam kriteria pengoptimalan yang kuat.

Pengoptimalan risiko yang kuat secara distribusi dapat mengakomodasi pengaturan yang bermusuhan dan tingkat kebisingan yang tinggi untuk merancang algoritme yang adil pada kelompok laten yang tidak diketahui. Meminimalkan risiko yang diharapkan dalam pembelajaran mesin dapat menghasilkan model dengan performa yang sangat buruk pada masukan terburuk. Uesato dkk. mendefinisikan tujuan pengganti yang dapat dilakukan terhadap risiko manipulasi sebenarnya yang cenderung sulit dilakukan secara komputasi. Mengoptimalkan risiko manipulasi memotivasi studi tentang kinerja model pembelajaran mesin pada input kasus terburuk. Risiko manipulasi seperti ini dapat diterapkan dalam situasi berisiko tinggi yang melibatkan sistem pembelajaran mesin untuk deteksi malware, visi komputer, robotika, pemrosesan bahasa alami, dan pembelajaran penguatan.

Wong dkk. mengusulkan pengklasifikasi berbasis ReLU yang kuat terhadap gangguan manipulasi yang dibatasi norma pada data pelatihan. Prosedur pengoptimalan yang kuat meminimalkan kerugian kasus terburuk pada perkiraan luar cembung dari kumpulan aktivasi lapisan akhir yang dicapai oleh gangguan yang dibatasi norma pada masukan. Masalah ini diselesaikan dengan program linier yang direpresentasikan sebagai jaringan saraf dalam yang dilatih dengan propagasi mundur kesalahan. Prediksi kelas dari pengklasifikasi kuat tersebut

terbukti tidak berubah dalam batas luar cembung ke nilai fungsi kerugian yang disebut "*adversarial polytope*." Analisis kerugian jaringan neural dalam kasus terburuk seperti itu berlaku bahkan untuk jaringan mendalam termasuk lapisan representasi seperti lapisan konvolusional.

Penelitian ini merupakan upaya untuk mendapatkan batas ketahanan yang dapat ditelusuri untuk wilayah gangguan yang merugikan di seluruh lapisan dalam jaringan yang dalam. Hal ini berbeda dengan penelitian tentang pemecah kombinatorial untuk memverifikasi properti jaringan saraf. Mereka termasuk pemecah teori modulo kepuasan (SMT) dan pendekatan pemrograman bilangan bulat. Namun, prosedur verifikasi yang canggih terlalu mahal secara komputasi untuk dapat diintegrasikan dengan mudah ke dalam prosedur pelatihan canggih yang ada saat ini. Tugas analisis data untuk memecahkan masalah pengoptimalan cembung yang kuat adalah dengan memecahkan masalah pengoptimalan di mana beberapa data masalah tidak diketahui tetapi termasuk dalam himpunan yang dibatasi.

Batasan ketahanan yang dapat dibuktikan pada kesalahan manipulasi dan hilangnya pengklasifikasi berasal dari solusi ganda dari masalah optimasi. Mereka dapat digunakan dalam definisi metrik kinerja yang dapat dibuktikan yang mengukur ketahanan dan deteksi serangan manipulasi dalam fungsi kerugian khusus yang dievaluasi pada kumpulan data pelatihan, pengujian, dan validasi. Sinha dkk. mengusulkan masalah optimasi yang kuat secara distribusi berdasarkan metrik jarak Wasserstein. Hal ini dapat diperluas ke dalam prosedur pelatihan adversarial pada pembaruan parameter model pembelajaran mesin yang dihadapkan pada gangguan terburuk pada data pelatihan.

Hal ini mampu mencapai ketahanan yang dapat dibuktikan untuk memuluskan fungsi kerugian dengan sedikit biaya manipulasi dibandingkan dengan minimalisasi risiko empiris dari kerugian pembelajaran. Hal ini dapat digunakan untuk memberikan jaminan sertifikasi atas kinerja komputasi dan statistik dari prosedur pelatihan manipulasi. Contoh manipulasi terbentuk karena gangguan fungsi kerugian halus Lagrangian dalam kasus terburuk. Pendekatan yang diusulkan untuk ketahanan distribusi terkait dengan model optimasi parametrik yang dibatasi pada momen, dukungan, dan penyimpangan arah dalam distribusi data pelatihan. Hal ini juga terkait dengan ukuran non-parametrik untuk jarak antar distribusi probabilitas seperti f-divergence, Kullback-Leibler divergence, dan Wasserstein distance.

Rauber dkk. membuat paket Python untuk menghasilkan gangguan manipulasi dan membandingkan kekokohan model pembelajaran mesin. Paket Python memiliki modul untuk membuat model pada data masukan, membuat prediksi keluaran sebagai probabilitas kelas, kriteria kesalahan klasifikasi untuk menentukan contoh manipulasi, ukuran jarak pada ukuran gangguan manipulasi, dan algoritma serangan untuk menghasilkan gangguan manipulasi yang diberikan. masukan, label, model, dan kriteria manipulasi. Algoritme serangan melakukan penyetelan hyperparameter untuk menemukan gangguan minimum.

Pembelajaran Generatif

Teman baik dkk. merangkum perlunya regularisasi dalam pembelajaran mendalam. Regularisasi dalam deep learning dibahas dengan mengacu pada underfitting, overfitting, bias, variance, dan generalisasi untuk mengontrol kompleksitas komputasi model machine

learning. Model pembelajaran mesin yang diregulasi berperforma baik tidak hanya pada data pelatihan tetapi juga pada masukan baru. Istilah regularisasi dalam fungsi tujuan pelatihan adalah penalti dan batasan yang dirancang untuk mengimbangi pengurangan kesalahan pengujian dengan kemungkinan peningkatan kesalahan pelatihan. Representasi renggang, ketahanan kebisingan, augmentasi kumpulan data, pelatihan adversarial, pembelajaran semi-supervisi, pembelajaran multitask, dan pembelajaran berjenis-jenis dicantumkan sebagai beberapa teknik regularisasi baru untuk memperkenalkan regularisasi ke dalam kerugian pembelajaran adversarial.

Teman baik dkk. juga mensurvei penggunaan optimasi analitis yang dikhususkan untuk meningkatkan prosedur pelatihan dalam pembelajaran mendalam. Mereka menggunakan pengoptimalan berbasis gradien sebagai dasar perbandingan dalam eksperimen pembandingan. Tujuannya adalah untuk menemukan parameter jaringan neural yang mengurangi fungsi biaya yang melibatkan ukuran kinerja yang dievaluasi pada kumpulan data pelatihan, ketentuan regularisasi yang dievaluasi pada kumpulan data pelatihan, dan kerugian manipulasi yang dievaluasi pada kumpulan data validasi. Di sini algoritma optimasi harus bersaing dengan strategi inisialisasi parameter, kecepatan pembelajaran adaptif selama pelatihan, dan informasi yang terkandung dalam turunan kedua dari fungsi biaya. Tujuan dari algoritma pembelajaran mesin yang dioptimalkan dapat dikatakan untuk meminimalkan kesalahan generalisasi yang diharapkan dengan menghitung rata-rata kesalahan pelatihan yang disebut risiko empiris. Minimalkan risiko empiris cenderung disesuaikan dengan kumpulan data pelatihan.

Hal ini harus diperhitungkan dalam kriteria konvergensi untuk pengoptimalan yang mengarah ke algoritme pengoptimalan batch, inkremental, stokastik, online, deterministik, dan acak untuk pembelajaran mendalam pada aliran data dinamis. Masalah yang tidak terkondisi, minimum lokal, titik pelana, gradien yang meledak, dan gradien yang tidak tepat dicantumkan sebagai beberapa tantangan teoretis dalam desain algoritma optimasi. Paradigma pembelajaran mesin seperti pembelajaran kurikulum, pembelajaran generatif, pembelajaran metrik, dan pembelajaran transfer berguna untuk memecahkan masalah tersebut dengan arsitektur jaringan saraf khusus.

Wang dkk. membahas hubungan antara ketahanan dan optimalisasi pembelajaran mendalam yang aman. Pelatihan manipulasi dengan serangan proyeksi penurunan gradien (PGD) dipilih sebagai masalah optimasi minmax. Masalah maksimalisasi bagian dalam menghasilkan contoh manipulasi dengan memaksimalkan kerugian klasifikasi. Minimisasi luar menghitung parameter model dengan meminimalkan kerugian manipulasi pada contoh manipulasi. Kondisi stasioner orde pertama (FOSC) yang memiliki solusi bentuk tertutup untuk optimasi terbatas diusulkan sebagai kerugian adversarial. Ini membangun strategi pelatihan dinamis untuk pembelajaran yang kuat dengan peningkatan bertahap dalam kualitas konvergensi dari contoh-contoh manipulasi yang dihasilkan.

Sebagai mekanisme pertahanan, teknik pelatihan manipulasi sebanding dengan teknik yang cukup kuat seperti penolakan masukan, regularisasi gradien, regularisasi Lipschitz, distilasi defensif, kompresi model, dan pelatihan manipulasi kurikulum. Skenario

serangan PGD yang dipilih sebanding dengan metode tanda gradien cepat (FGSM), serangan peta saliency berbasis Jacobian (JSMA), serangan C&W, dan serangan berbasis Frank-Wolfe. Meningkatkan kekerasan komputasi contoh adversarial secara bertahap adalah ide yang didasarkan pada paradigma pembelajaran kurikulum untuk pembelajaran mesin. Hal ini mempercepat konvergensi dan meningkatkan generalisasi jaringan pembelajaran mendalam. Kurikulum pembelajaran dalam mekanisme pengurutan berurutan dirancang untuk pelatihan manipulasi. Hasil eksperimen dari pelatihan manipulasi ini dibandingkan dengan serangan canggih terhadap WideResNet.

Model Generatif Mendalam untuk Pembelajaran Tantangan Teori Game

Kunin dkk. mempelajari lanskap kerugian dalam autoencoder linier teregulasi (LAE) yang bertindak sebagai model untuk pembelajaran representasi mendalam. Autoencoder dilatih untuk meminimalkan jarak antara data dan rekonstruksinya. Mereka mempelajari subruang yang direntang oleh vektor dasar yang dipelajari dari data pelatihan. LAE terhubung dengan model regresi penurunan peringkat seperti analisis komponen utama (PCA). Efek regularisasi L2 pada pola ortogonalitas di encoder dan decoder diselidiki. LAE diartikan sebagai proses generatif. Autoencoder denoising dan autoencoder kontraktif dibahas sebagai varian LAE.

Vincent dkk. menumpuk lapisan autoencoder denoising di jaringan saraf dalam yang mampu menunjukkan kesalahan klasifikasi yang lebih rendah. Representasi data pelatihan tingkat tinggi diperoleh dari kriteria denoising yang bertindak sebagai tujuan tanpa pengawasan untuk pendeteksi fitur. Mampu mendongkrak kinerja mesin vektor pendukung dalam klasifikasi multi-label. Kesalahan rekonstruksi dalam autoencoder denoising yang diusulkan dapat dianggap sebagai peningkatan estimasi kemungkinan log pada mesin Boltzmann terbatas stokastik yang menggunakan pembaruan divergensi kontras. Prinsip infomax dari analisis komponen independen dimanfaatkan sebagai kriteria denoising yang memaksimalkan informasi timbal balik antara variabel acak masukan dan representasi tingkat yang lebih tinggi.

Rata-rata empiris informasi timbal balik pada sampel pelatihan diambil sebagai penduga yang tidak bias untuk pembelajaran tanpa pengawasan. Fungsi kerugian encoder dinyatakan sebagai fungsi affine+sigmoid, sedangkan fungsi kerugian decoder dinyatakan sebagai affine dengan kehilangan kesalahan kuadrat atau affine+sigmoid dengan kerugian lintas entropi. Meminimalkan kesalahan rekonstruksi menggabungkan fungsi kehilangan encoder dan decoder saat melatih autoencoder bertumpuk. Kesalahan rekonstruksi seperti itu setara dengan memaksimalkan batas bawah informasi timbal balik antara masukan dan representasi yang dipelajari. Representasi pembelajaran berdimensi lebih rendah dapat dianggap sebagai representasi input terkompresi yang lossy.

Representasi yang dipelajari dapat berupa representasi pengkodean yang jarang, representasi terkompresi yang padat, dan representasi ukuran variabel yang cocok untuk mengekstraksi fitur-fitur yang berguna dalam konstruksi pengklasifikasi jaringan saraf dalam. Kriteria denoising teratur tanpa pengawasan yang berhasil menentukan dan mempelajari keragaman data. Contoh manipulasi cenderung lebih jauh dari keragaman data dibandingkan

contoh pelatihan. Proses kerusakan data dapat diparameterisasi menjadi sinyal pelatihan berbasis denoising yang diberikan ke algoritma pembelajaran autoencoder denoising mendalam.

Kingma dkk. memperkenalkan inferensi variasi stokastik untuk pembelajaran yang efisien dalam model probabilistik terarah untuk autoencoder. Model inferensi perkiraan diusulkan untuk i.i.d. kumpulan data yang mengarah ke distribusi posterior yang sulit diselesaikan dalam estimasi kemungkinan maksimum dan inferensi maksimum a posteriori di mana algoritma pemaksimalan ekspektasi tidak dapat digunakan. Model pengenalan pola dalam pembuat enkode probabilistik dioptimalkan untuk melakukan perkiraan inferensi posterior yang efisien tanpa menggunakan metode pengambilan sampel yang mahal. Ini memiliki aplikasi dalam pengenalan pembelajaran manipulasi, denoising, inpainting, representasi, dan visualisasi dalam lingkungan online dan non-stasioner. Autoencoder yang diusulkan meniru proses acak tersembunyi yang mendasari distribusi data pelatihan untuk menghasilkan data variasi buatan yang menyerupai data pelatihan yang diamati. Hal ini dapat diperluas ke arsitektur generatif yang mendalam, jaringan Bayesian yang dinamis, dan pembelajaran yang diawasi dengan variabel laten dan distribusi kebisingan yang rumit.

Inferensi Variasi Alain dkk. menyarankan agar autoencoder mempelajari struktur manifold lokal dari distribusi data yang mendasari data pelatihan. Jadi fungsi rekonstruksi yang diatur dalam autoencoder mampu mengkarakterisasi bentuk fungsi kepadatan probabilitas yang menghasilkan data sebagai bidang vektor di sekitar manifold. Autoencoder menangkap turunan kepadatan log sehubungan dengan masukan sebagai fungsi pencocokan skor denoising. Fungsi pencocokan skor muncul dari trade-off antara meminimalkan kesalahan rekonstruksi dan mengatur autoencoder. Kriteria pelatihan autoencoder dianggap sebagai alternatif yang mudah untuk memperkirakan kemungkinan maksimum. Autoencoder dapat bertindak sebagai model kepadatan implisit untuk mengambil sampel esensi distribusi data target yang mendasari data pelatihan.

Dalam model saraf generatif, Mondal et al. berhipotesis bahwa dimensi ruang laten autoencoder telah mempengaruhi kualitas data yang dihasilkan. Kualitas data yang dihasilkan dibandingkan antara model autoencoder dan jaringan manipulasi generatif. Performa optimal diperoleh ketika dimensi ruang laten autoencoder cocok dengan dimensi ruang laten generatif. Mask Adversarial Auto-Encoder (MaskAAE) diusulkan untuk memenuhi kondisi tersebut dengan menutupi dimensi laten palsu. Oleh karena itu, metode variasi mendalam dapat dianggap sebagai model generatif probabilistik. Representasi data ruang laten disebabkan oleh encoder deterministik atau stokastik. Data yang dihasilkan berasal dari dekoder yang mewujudkan kumpulan aproksimasi fungsi yang dapat dipelajari. Distribusi data dalam ruang laten mengikuti distribusi probabilitas yang diketahui sehingga pengambilan sampel dapat dilakukan. Prosedur penyembunyian algoritmik meminimalkan kesalahan rekonstruksi berbasis norma dan metrik divergensi seperti divergensi JS, divergensi KL, atau jarak Wasserstein antara distribusi kepadatan campuran sebelumnya dan distribusi laten yang dikodekan bertopeng.

Zhao dkk. melatih model variabel laten yang dalam untuk struktur diskrit seperti urutan teks dan gambar terdiskritisasi dalam transfer gaya tekstual. Mereka adalah perpanjangan dari kerangka autoencoder Wasserstein dan memformalkan masalah optimasi autoencoder sebagai masalah transportasi yang optimal. Distribusi sebelumnya yang telah diperbaiki dan dipelajari yang berbeda dari generator berparameter dalam autoencoder yang diatur secara berlawanan dapat menargetkan representasi generatif di ruang keluaran. Generator parametrik berbasis pembelajaran transfer dilatih untuk mengabaikan atribut input yang ditargetkan. Ini dapat digunakan untuk transfer sentimen atau gaya antara domain sumber dan target yang tidak selaras.

Manipulasi gambar dan kalimat dapat dilakukan di ruang laten melalui interpolasi dan aritmatika vektor untuk menyebabkan perubahan pada ruang keluaran. Membangun gaya interpolasi memerlukan pencarian kombinatorial. Pengklasifikasi atribut ruang laten diperkenalkan untuk melatih pembuat enkode secara berlawanan. Autoencoder semacam itu mengakomodasi transformasi halus dalam ruang laten kontinu yang diatur secara berlawanan untuk menghasilkan modifikasi kompleks dari keluaran yang dihasilkan dalam manifold data. Ukuran divergensi informasi seperti f-divergensi atau jarak Wasserstein meminimalkan perbedaan antara distribusi kode yang dipelajari dari distribusi sebenarnya dan distribusi model. Hilangnya entropi silang di autoencoder membatasi total jarak variasi antara model/distribusi data.

Decoder diskrit seperti jaringan saraf berulang dapat dimasukkan ke dalam distribusi model. Di sini fungsi tujuan yang tidak dapat dibedakan diselesaikan dengan metode gradien kebijakan dalam pembelajaran penguatan dan distribusi Gumbel-Softmax untuk memperkirakan pengambilan sampel data diskrit. Pembelajaran autoencoder dapat diartikan sebagai mempelajari model generatif yang mendalam dengan variabel laten selama ruang encode yang terpinggirkan sama dengan sebelumnya. Regularisasi manipulasi berdampak pada pengkodean diskrit, kelancaran encoder, rekonstruksi dalam decoder, dan manipulasi output melalui prior. Model variabel laten dalam yang dihasilkan sensitif terhadap pengaturan pelatihan dan ukuran kinerja. Meningkatkan ketahanan mereka terhadap persaingan akan menghasilkan model untuk struktur terpisah yang kompleks seperti dokumen.

Mescheder dkk. menyatukan autoencoder variasional (VAE) dan jaringan manipulasi generatif (GAN). VAE dinyatakan sebagai model variabel laten untuk mempelajari distribusi probabilitas kompleks dari data pelatihan. Perpanjangan yang disebut adversarial variasional Bayes (AVB) dengan model inferensi yang dapat ditafsirkan telah diusulkan. Ia memiliki jaringan diskriminatif tambahan yang merumuskan estimasi kemungkinan maksimum sebagai permainan dua pemain seperti permainan di GAN. Model generatif mendalam yang diusulkan lebih baik daripada model generatif seperti Pixel-RNN, PixelCNN, NVP nyata, dan jaringan generatif Plug & Play. Dalam estimasi kemungkinan log, GAN memiliki keunggulan dalam menghasilkan representasi generatif dari data pelatihan serta VAE untuk menghasilkan model generatif dan model inferensi.

Di sini model inferensi yang sangat ekspresif dikombinasikan dengan dekoder yang kuat memungkinkan VAE memanfaatkan representasi ruang laten untuk mencapai kesalahan rekonstruksi. Kerugian manipulasi dalam model inferensi mendorong agregat posterior menjadi dekat dengan prior dibandingkan variabel laten. Estimasi parameter Bayesian mendekati distribusi posterior sebagai model probabilistik. Hal ini dapat memperkirakan batas bawah variasional untuk mempelajari model variabel laten yang meminimalkan perbedaan KL antara distribusi data pelatihan dan data laten. Model probabilistik mampu mempelajari distribusi posterior multimodal dan menghasilkan sampel untuk kumpulan data yang kompleks. Jaringan konvolusional dalam digunakan sebagai jaringan dekoder.

Arsitektur jaringan encoder yang terdiri dari vektor kebisingan berbasis pembelajaran mampu menghitung momen distribusi data laten secara efisien yang dikondisikan pada distribusi data masukan. Model inferensi dapat mewakili kelompok distribusi kondisional apa pun atas variabel laten. Validasi eksperimental dibandingkan dengan metode annealed important sampling (AIS) untuk model generatif berbasis decoder. Blei dkk. membahas pemanfaatan inferensi variasi dan optimasi dalam statistik Bayesian untuk memperkirakan kepadatan probabilitas posterior yang mahal secara komputasi. Metode variasi ditemukan lebih cepat dibandingkan metode pengambilan sampel seperti rantai Markov Monte Carlo. Mereka mengukur perbedaan informasi antara perkiraan distribusi data dengan kelompok kepadatan target yang diajukan. Di sini inferensi variasi bidang rata-rata berlaku untuk model keluarga eksponensial seperti model entropi maksimum yang membentuk fungsi kerugian dalam pembelajaran mesin.

Mereka dapat digunakan dalam optimalisasi stokastik pembelajaran manipulasi teoritis permainan. Grunwald dkk. menunjukkan teori kesetaraan antara memaksimalkan entropi relatif umum dan meminimalkan kemungkinan kerugian terburuk yang didasarkan pada permainan zero-sum antara pengambil keputusan dan Alam. Tindakan Bayes yang kuat ditemukan untuk meminimalkan perbedaan atau perbedaan antara distribusi sehingga memaksimalkan entropi. Mereka dinyatakan sebagai solusi terhadap teorema minmax pada divergensi Kullback-Leibler yang dihitung untuk kelompok kepadatan target eksponensial umum. Teorema minmax disebut teorema kapasitas redundansi dalam teori informasi. Entropi relatif umum adalah fungsi ketidakpastian yang terkait dengan fungsi kerugian untuk melatih model pembelajaran mesin.

Model aditif untuk inferensi statistik yang didasarkan pada divergensi Bregman adalah kasus khusus dari keluarga eksponensial umum. Mereka dapat digunakan untuk mendapatkan aturan penilaian seperti skor Brier dan skor Bregman dalam permasalahan pengambilan keputusan untuk klasifikasi multi-label. Properti Pythagoras dari divergensi Kullback-Leibler mengarah pada interpretasi inferensi entropi relatif minimum sebagai operasi proyeksi informasi antara distribusi data manipulasi dan pelatihan pada ruang sampel diskrit. Hal ini dapat diperluas dengan masalah optimasi terkait entropi berdasarkan teori informasi tentang ketidaksetaraan momen dan keluarga entropi umum. Entropi umum tersebut mencakup entropi Renyi dan informasi Fisher yang diinterpretasikan dari perspektif minimax.

Autoencoder Musuh Bengio dkk. membahas autoencoder denoising dan kontraktif untuk fungsi kepadatan yang dipelajari secara implisit yang memperkirakan distribusi penghasil data yang mendasarinya. Ia dapat menangani fitur-fitur yang bernilai diskrit dan berkelanjutan dengan kerusakan yang sewenang-wenang. Kerugian akibat rekonstruksi dilihat sebagai estimasi log-likelihood. Regularisasi rekonstruksi mencegah autoencoder mempelajari fungsi identitas sederhana dan malah berperilaku sebagai pembelajar fitur untuk pembelajaran yang diawasi. Dalam interpretasi probabilistik dari kerugian rekonstruksi autoencoder, kesalahan rekonstruksi denoising memperkirakan fungsi energi untuk pencocokan skor dalam mesin Boltzmann yang dibatasi Gaussian.

Mnih dkk. mengusulkan metode inferensi perkiraan non-iteratif untuk melatih jaringan kepercayaan sigmoid. Ini menerapkan pengambilan sampel tepat yang efisien dari posterior variasional dalam jaringan feedforward. Algoritme pelatihan memperbarui model jaringan saraf dan jaringan inferensi dengan memaksimalkan batas bawah variasi pada kemungkinan log marjinal. Hasil eksperimen terbukti lebih baik dibandingkan algoritma bangun-tidur untuk pelatihan stokastik. Beberapa garis dasar dieksplorasi untuk pengurangan varians dalam jaringan inferensi yang berhubungan dengan decoder. Tujuan variasional tersebut dapat digunakan untuk melatih pembuat enkode probabilistik dalam pembelajaran manipulasi. Mereka dirancang dalam kerangka teori informasi seperti panjang deskripsi minimum untuk pengkodean distribusi data non-stasioner.

Contoh manipulasi menyebabkan peningkatan kesalahan generalisasi dan waktu inferensi jaringan pembelajaran mendalam. Kyatham dkk. mengusulkan mekanisme pertahanan terhadap contoh-contoh manipulasi berdasarkan model generatif ruang laten yang diatur. Ini melibatkan filter manipulasi yang mengkodekan ruang laten terkuantisasi dari data yang banyak tunduk pada manipulasi. Filter manipulasi tidak dapat diakses oleh musuh atau pengklasifikasi. Ia memiliki mekanisme inferensi variasional dalam ruang laten generatif yang teratur, terkuantisasi, untuk memetakan kembali data manipulasi yang dikodekan ke manifold data pelatihan yang sebenarnya.

Kekuatan manipulasi dari data yang didekode ditunjukkan dalam berbagai skenario serangan yang melibatkan metode white-box dan blackbox. Oleh karena itu, autoencoder variasional dapat digunakan untuk menjelajahi subruang fitur yang terdiri dari contoh-contoh manipulasi. Subruang fitur tersebut dapat dimasukkan ke dalam pelatihan ulang manipulasi yang kuat terhadap musuh tingkat pertama tetapi tidak mampu bertahan melawan serangan kotak hitam. Serangan blackbox juga dapat dioptimalkan untuk menghindari ketahanan terhadap gradien yang dikaburkan yang diperoleh dari perkiraan turunan dalam perkiraan fungsi, parameterisasi ulang, dan penghitungan ekspektasinya.

Encoder ruang laten yang diusulkan menjaga jarak antar sampel di bawah transformasi ruang metrik dari data ke manifold laten. Encoder digunakan dalam model generatif terkuantisasi yang memungkinkan eksplorasi stokastik pada lingkungan besar di ruang laten. Kode laten kemudian dipetakan kembali ke data yang sah. Dengan demikian, decoder dapat digunakan untuk mengubah contoh-contoh yang bersifat manipulasi menjadi sampel-sampel yang mendekati non-manipulasi. Model inferensi yang diusulkan disebut

autoencoder variasi terkuantisasi terbatas Lipschitz (LQ-VAE). Dengan kuantisasi biner sederhana dari ruang laten, detektor manipulasi dapat digunakan.

Gulrajani dkk. menyajikan PixelVAE yang memiliki dekoder autoregresif berdasarkan PixelCNN. PixelVAE mampu mempelajari representasi laten yang berguna untuk pemodelan gambar alami dengan detail halus. PixelVAE dapat diperluas untuk memiliki beberapa lapisan stokastik untuk memodelkan tidak hanya piksel keluaran tetapi juga peta fitur laten tingkat yang lebih tinggi. Kemungkinan kondisional autoregresif dieksplorasi dalam konteks aplikasi analisis data seperti pemodelan kalimat. Distribusi keluaran untuk jaringan generatif dan inferensi dapat didekomposisi dan difaktorkan berdasarkan variabel laten untuk mendapatkan kemungkinan log untuk data yang direkonstruksi yang diatur oleh divergensi KL dari perkiraan posterior terhadap laten dengan prior autoregresif. Representasi laten dari data masukan dapat diterapkan pada pembelajaran representasi mendalam dalam klasifikasi semi-supervisi.

Hou dkk. mengusulkan fungsi kerugian untuk VAE yang menerapkan konsistensi fitur mendalam yang menjaga karakteristik korelasi spasial dari masukan untuk memberikan kualitas persepsi yang lebih baik. Fitur tersembunyi dari jaringan neural konvolusional dalam (CNN) yang telah dilatih sebelumnya menentukan hilangnya persepsi fitur untuk pelatihan VAE. Alih-alih merekonstruksi pengukuran piksel demi piksel, kehilangan persepsi fitur menentukan perbedaan antara representasi gambar tersembunyi yang telah diekstraksi dari CNN mendalam yang telah dilatih sebelumnya seperti AlexNet, VGGNet, dan ImageNet. Vektor laten yang diperoleh dari VAR mencapai kinerja canggih dalam prediksi atribut wajah.

Distribusi vektor laten dapat dikontrol berdasarkan divergensi KL dari variabel acak Gaussian. Hal ini dikombinasikan dengan kerugian rekonstruksi untuk melatih VAE. Kemudian VAE mendalam yang dikondisikan atribut seperti penulis perhatian berulang yang mendalam (DRAW) dapat diperluas ke pembelajaran semi-supervisi dengan label kelas yang menggabungkan mekanisme perhatian dengan kerangka pengkodean otomatis variasional sekuensial. Kinerja VAE juga dapat ditingkatkan dengan regularisasi diskriminatif atas kerugian rekonstruksi yang dicapai oleh diskriminator GAN pada representasi fitur yang dipelajari di VAE. Hilangnya persepsi fitur dapat ditentukan oleh transfer gaya saraf dan skor klasifikasi pada fitur individual dari CNN mendalam yang telah dilatih sebelumnya.

Hou dkk. memperluas VAE yang konsisten dengan fitur mendalam untuk mengimplementasikan mekanisme pelatihan manipulasi generatif konvolusional mendalam yang mempelajari penyematan fitur dalam manipulasi atribut wajah. Strategi ekstraksi fitur multiview kemudian diusulkan untuk mengekstraksi representasi gambar efektif yang berguna dalam tugas prediksi atribut wajah. Model generatif untuk database gambar berguna untuk menghasilkan gambar realistis dari input acak, mengompresi database menjadi parameter model yang dipelajari, dan mempelajari representasi data tak berlabel yang dapat digunakan kembali yang dapat diterapkan ke dalam tugas pembelajaran yang diawasi seperti klasifikasi gambar. Diskriminator yang diusulkan menyeimbangkan keluaran antara kerugian rekonstruksi gambar dan kerugian manipulasi. VAE yang diusulkan dapat mempelajari informasi semantik atribut wajah secara linier dalam ruang laten yang dipelajari. Itu dapat

mengekstrak representasi atribut wajah yang diskriminatif. Gambar dapat ditransformasikan antar kelas dengan kombinasi linier sederhana dari vektor latennya. Fitur khusus atribut dapat dikodekan untuk gambar beranotasi guna memanipulasi atribut terkait dari gambar tertentu sambil memperbaiki atribut lainnya. Dengan demikian pelatihan adversarial yang diusulkan dalam VAE dapat dikondisikan pada label kelas dan atribut visual yang diperoleh dari data manifold gambar alam.

Larsen dkk. menyajikan autoencoder yang dapat mengukur kesamaan dalam ruang data berdasarkan representasi fitur yang dipelajari. Representasinya diperoleh dengan menggabungkan autoencoder variasional (VAE) dengan jaringan manipulasi generatif (GAN) ke dalam model generatif tanpa pengawasan. Dalam distribusi data yang direkonstruksi, kesalahan berdasarkan elemen diganti dengan kesalahan berdasarkan fitur yang menawarkan invarian terhadap terjemahan gambar. Hasil pelatihan menghasilkan representasi gambar laten dengan faktor variasi yang terurai. Fitur visual tingkat tinggi dapat dimodifikasi menggunakan aritmatika vektor. Dekoder VAE dan generator GAN berbagi parameter yang dilatih bersama. Oleh karena itu, model generatif dapat ditingkatkan dengan mengukur kesamaan yang dipelajari yang dimasukkan ke dalam metrik kualitas rekonstruksi pada kelas objek. Pada saat yang sama, diskriminator GAN dapat digunakan untuk mengukur kesamaan sampel. Metode yang diusulkan juga dapat diartikan sebagai GAN mempelajari distribusi data yang kompleks dengan prior yang dibatasi oleh VAE.

Tran dkk. mengusulkan formulasi di mana sampel yang direkonstruksi dari autoencoder (AE) dimasukkan sebagai sampel "nyata" untuk diskriminator dalam GAN. Hal ini mempengaruhi kriteria konvergensi GAN. Selanjutnya, batasan jarak data laten diterapkan pada jaringan encoder. Ini meminimalkan jarak antara sampel laten dan sampel data. Batasan jarak skor diskriminator menyelaraskan distribusi sampel yang dihasilkan dengan sampel data sebenarnya. Kedua batasan tersebut memandu generator Dist-GAN yang diusulkan untuk mensintesis sampel yang mirip dengan distribusi data pelatihan. Oleh karena itu, proses pelatihan manipulasi di GAN dapat dikombinasikan dengan AE untuk menghasilkan sampel distribusi data tanpa memperkirakannya secara eksplisit. Pengurangan dimensi dalam AE dapat digunakan untuk menyeimbangkan kapasitas diskriminator dan generator yang menyebabkan masalah konvergensi seperti hilangnya gradien dan keruntuhan mode. Skor kompetitif yang dihasilkan oleh diskriminator dalam Dist-GAN yang stabil dapat digunakan untuk menemukan perbedaan multimodal antara distribusi data adversarial dan pelatihan. Ukuran divergensi informasi seperti divergensi KL dan divergensi JS dapat digabungkan dengan hilangnya rekonstruksi AE untuk melatih model inferensi empiris. Di sini, kerugian manipulasi teoritis permainan dapat diartikan sebagai istilah regularisasi. Menggabungkan AE dengan diskriminator di GAN memungkinkan kita mempelajari kondisi konvergensi pembelajaran adversarial teoretis permainan dari perspektif tugas visi komputer.

Makhzani dkk. mengusulkan autoencoder probabilistik yang disebut adversarial autoencoder (AAE). Ia melakukan inferensi variasional dengan mencocokkan agregat posterior autoencoder dengan distribusi sebelumnya yang berubah-ubah yang bertindak sebagai istilah regularisasi. Seorang pembuat enkode belajar mengubah distribusi data ke

distribusi sebelumnya yang bertindak sebagai distribusi pengkodean. Decoder mempelajari model generatif mendalam yang memetakan distribusi yang diterapkan sebelum distribusi posterior, mencocokkan distribusi data asli dengan distribusi decoding. Decoder AAE mempelajari model generatif yang mendalam seperti jaringan pencocokan momen generatif dan jaringan manipulasi generatif. Jaringan pengenalan dapat memprediksi distribusi decoding posterior terhadap variabel laten.

Autoencoder dilatih dengan kriteria kesalahan rekonstruksi antara distribusi model dan distribusi data. Hal ini dikombinasikan dengan kriteria pelatihan manipulasi yang secara diskriminatif memprediksi apakah sampel yang dihasilkan muncul dari kode tersembunyi autoencoder atau dari distribusi sampel yang ditentukan oleh pengguna. Prosedur pelatihan adversarial terdiri dari fase rekonstruksi untuk melatih autoencoder diikuti dengan fase regularisasi untuk memperlancar jaringan adversarial. Fase rekonstruksi memperbarui encoder dan decoder untuk meminimalkan kesalahan rekonstruksi input. Fase regularisasi memperbarui jaringan diskriminatif untuk membedakan antara sampel sebenarnya yang dihasilkan menggunakan sampel sebelumnya dan sampel yang dihasilkan yang merupakan kode tersembunyi yang dihitung oleh autoencoder. Kemudian generator diperbarui untuk membingungkan jaringan diskriminatif. Generator jaringan manipulasi juga merupakan encoder dari autoencoder.

Setelah pelatihan, dekoder autoencoder menentukan model generatif yang memetakan penerapan sebelum distribusi data. Encoder dapat berupa salah satu fungsi deterministik, distribusi stokastik seperti posterior Gaussian, dan aproksimator universal posterior yang menggabungkan distribusi data pelatihan dan manipulasi. Trik parametrisasi ulang digunakan dalam propagasi balik kesalahan melalui encoder distribusi stokastik. Dalam kasus pendekatan universal posterior, prosedur pelatihan manipulasi ditafsirkan sebagai metode pengambilan sampel yang efisien dari agregat posterior. Prior yang diberlakukan dapat berupa distribusi yang rumit dalam ruang berdimensi tinggi seperti distribusi swiss roll tanpa bentuk fungsional eksplisit untuk distribusi tersebut.

Fase rekonstruksi pelatihan manipulasi juga dapat memasukkan informasi campuran label kelas untuk membentuk distribusi kode tersembunyi dengan lebih baik. Di sini, pengklasifikasi semi-supervisi meminimalkan biaya entropi silang yang dihitung pada estimasi posterior bersyarat untuk setiap mini-batch berlabel. Desain AAE menunjukkan bahwa model generatif mendalam dapat dilatih secara berlawanan tidak hanya dengan metode pengambilan sampel seperti mesin Boltzmann yang dibatasi, tetapi juga metode variasi seperti autoencoder berbobot penting. AAE yang diusulkan terbukti memiliki aplikasi dalam klasifikasi semi-supervisi, pengelompokan tanpa pengawasan, reduksi dimensi, dan visualisasi data.

Scutari dkk. menganalisis pemodelan teori permainan sebagai sekumpulan masalah optimasi cembung berpasangan dalam matematika terapan. Masalah optimasi cembung seperti itu dipelajari secara luas dalam pemrosesan sinyal untuk desain sistem komunikasi pengguna tunggal dan multipengguna. Di sini, pendekatan teori permainan kooperatif dan non-kooperatif dapat digunakan untuk memodelkan keseimbangan dalam masalah

komunikasi dan jaringan. Optimalisasi tersebut juga dapat digeneralisasikan ke masalah ketimpangan variasi dalam analisis non-linier.

Dengan demikian pemrosesan sinyal dapat digunakan dalam studi tentang keberadaan dan keunikan keseimbangan Nash dalam pembelajaran adversarial teori permainan. Selanjutnya, algoritma komputasi terdistribusi berulang dapat dirancang untuk mempelajari sifat konvergensi dan pemrograman keseimbangan dari pemodelan teoritis permainan. Aplikasi pembelajaran manipulasi terkait juga memiliki relevansi dalam pemrosesan sinyal dan aplikasi komunikasi seperti berbagi sumber daya dalam jaringan komunikasi multihop, jaringan radio kognitif, jaringan ad hoc nirkabel, dan jaringan kabel per-to-peer.

Gidel dkk. mengeksplorasi kerangka ketidaksetaraan variasional sebagai metode optimasi titik pelana untuk merancang pelatihan manipulasi. Pelatihan GAN diperluas untuk mencakup rata-rata ketidaksetaraan variasional dan ekstrapolasi. Dalam pemrograman matematika, masalah pertidaksamaan variasional menggeneralisasi kondisi stasioner untuk permainan dua pemain. Pada titik stasioner, turunan arah fungsi biaya adalah non-negatif dalam segala arah yang memungkinkan untuk optimasi. Mereka dapat digeneralisasikan ke bidang vektor kontinu. Masalah pertidaksamaan variasional menemukan himpunan optimal pada bidang vektor. Pemodelan teoritis permainan dalam pemodelan generatif mendalam dapat dieksplorasi dalam kerangka ketimpangan variasional untuk menghasilkan ketidaksetaraan variasional stokastik dengan batasan terbatas dan minimalisasi penyesalan dalam pembelajaran online. Di sini permainan bukan nol adalah tujuan pembelajaran GAN. Ketimpangan variasional dapat dimanfaatkan dalam berbagai algoritma optimasi praktis.

Harker dkk. mengulas masalah ketidaksetaraan variasional dimensi terbatas dalam teori permainan terutama untuk model non-linier. Pemecahan model kesetimbangan adalah topik yang disebut pemrograman kesetimbangan dalam optimasi non-linier. Mereka dapat digunakan untuk menghasilkan metode komputasi numerik untuk mempelajari sifat konvergensi keseimbangan teori permainan. Analisis sensitivitas dan stabilitas kesetimbangan terhadap perubahan parameter model merupakan bagian penting dari keberadaan dan keunikan solusi. Pemodelan numerik yang dihasilkan dapat diintegrasikan ke dalam pembelajaran manipulasi teoritis permainan untuk mengoptimalkan pemodelan dinamika dan komputasi dalam skenario serangan berulang dan mekanisme pertahanan. Daniele mengontekstualisasikan kembali pemodelan dinamika sebagai ketidaksetaraan variasional evolusioner dalam jaringan dinamis yang berkembang seiring waktu. Pemodelan dinamika mempunyai penerapan di bidang keuangan, ekonomi, ilmu komputer, dan matematika.

Fitur Pemulusan Distribusi Vincent dkk. memperluas pra-pelatihan tanpa pengawasan ke pembelajaran representasi tanpa pengawasan sehingga representasi yang dipelajari tahan terhadap kerusakan parsial pada pola masukan. Model generatif mendalam kemudian dibuat dengan menumpuk autoencoder denoising untuk berbagai pembelajaran. Representasi tingkat yang lebih tinggi dari pola yang diamati dihasilkan dengan mengoptimalkan kriteria pembelajaran lokal tanpa pengawasan. Kriteria pelatihan global kemudian diusulkan untuk

mengoptimalkan kinerja yang sesuai dengan tugas yang ada. Representasi tanpa pengawasan bertindak sebagai inisialisasi algoritma optimasi yang mampu menghindari solusi yang buruk. Mesin Boltzmann terbatas yang dilatih dengan divergensi kontras dan berbagai autoencoder ternyata mendapat manfaat dari proses pelatihan semacam itu. Perspektif teori informasi kemudian disediakan untuk menganalisis autoencoder tangguh yang secara efisien memodelkan distribusi data kompleks dan menunjukkan kinerja generalisasi yang unggul pada ketergantungan dalam distribusi data yang mencirikan masukan yang diamati.

Prosedur denoising yang diusulkan sebanding dengan augmentasi kumpulan data pelatihan dengan pola manipulasi. Namun prosedur denoising tidak menggunakan pengetahuan sebelumnya tentang topologi gambar dan label kelas untuk pembelajaran yang diawasi. Untuk menangani korupsi karena kebisingan yang bersifat manipulasi atau sebaliknya, prosedur denoising tidak menghasilkan fungsi yang mulus untuk regularisasi tetapi mempelajari informasi kekokohan dalam inferensi variasional atas penghancuran informasi yang besar dan non-aditif. Entropi silang rekonstruksi adalah tujuan pelatihan. Ini memaksimalkan batas bawah informasi timbal balik antara kerugian pelatihan dan kerugian manipulasi. Namun, model grafis terarah yang mendalam terus menimbulkan tantangan optimasi dalam pemodelan generatif yang mendalam terutama untuk mempelajari konsep tingkat tinggi dari input multimodal.

Zhao dkk. mengusulkan model generatif mendalam yang dapat belajar dari hierarki fitur dalam tugas pembelajaran yang diawasi. Beberapa lapisan variabel laten dilatih dengan metode variasional. Tidak seperti metode diskriminatif yang mempelajari hierarki fitur invarian dan lokal, metode variasional yang diusulkan mempelajari fitur hierarki yang dapat diinterpretasikan dengan melestarikan informasi pada kumpulan data gambar alami. Representasi yang dipelajari dapat digeneralisasikan ke model terlatih yang mendukung inferensi statistik. Lapisan tersembunyi dari variabel laten dikarakterisasi dalam dua desain. Desain pertama secara rekursif menumpuk model generatif dengan asumsi bahwa lapisan bawah saja berisi informasi untuk merekonstruksi distribusi data dan informasi tersebut tidak bergantung pada kelompok distribusi tertentu yang digunakan untuk menentukan hierarki.

Desain kedua berfokus pada model variabel laten satu lapis di mana fitur tingkat tinggi diposisikan pada bagian tertentu dari kode laten dan fitur tingkat rendah diposisikan pada bagian lain. Pendekatan ini disebut autoencoder tangga variasional. Ini memaksimalkan kemungkinan log marjinal pada kumpulan data pelatihan. Kemungkinannya rumit dan sulit untuk model generatif. Marginalisasi ini disebabkan oleh variabel laten autoencoder. Mengikuti model inferensi variasional, batas bawah bukti (ELBO) yang melibatkan divergensi Kullback-Leibler dioptimalkan sebagai solusi untuk optimasi kemungkinan marjinal yang sulit dilakukan. Inferensi seperti itu terbukti menghasilkan representasi terstruktur yang dipelajari yang lebih baik daripada mengasumsikan struktur independensi Markov dalam variabel laten untuk memfaktorkan distribusi inferensi menurut autoencoder variasi hierarki autoregresif.

Sønderby dkk. mengusulkan autoencoder variasi tangga untuk pembelajaran representasi fitur tanpa pengawasan. Ini secara rekursif mengoreksi distribusi generatif dengan perkiraan kemungkinan yang bergantung pada data. Kemungkinan log prediktif

memberikan batas bawah ke inferensi bottom-up dalam autoencoder variasional berlapis. Ini juga dapat digunakan dalam desain hierarki variabel laten yang terdistribusi secara mendalam dalam model pembelajaran inferensi dan generatif. Hierarki variabel stokastik bersyarat dalam VAE tersebut ditafsirkan sebagai representasi model terfaktor yang efisien secara komputasi. Mereka memperkirakan perkiraan variasional posterior yang membatasi posterior sejati yang sulit diatur.

Hal ini diperkirakan dengan pemodelan struktur ketergantungan antara inferensi kemungkinan bottom-up dan pemodelan informasi generatif top-down dalam pembelajaran mendalam. Parameterisasi VAE seperti itu memungkinkan interaksi antara sinyal bottom-up dan top-down seperti pada autoencoder tangga variasional. Performa generatif dari distribusi variasional dibandingkan dengan garis dasar VAE seperti proses Gaussian variasional, aliran normalisasi, autoencoder berbobot kepentingan, dan model generatif mendalam tambahan. Divergensi KL yang membatasi kriteria pelatihan log-likelihood diperkirakan menggunakan pengambilan sampel Monte Carlo. Algoritma propagasi mundur stokastik digunakan untuk mengoptimalkan parameter generatif dan inferensi.

Dalam inferensi VAE, setiap lapisan stokastik ditentukan sebagai distribusi Gaussian yang terfaktor sepenuhnya. Istilah regularisasi variasi dimasukkan ke dalam fungsi kerugian untuk penaksir distribusi kemungkinan log generatif. Model ini juga dapat mengakomodasi pembagian parameter secara eksplisit antara distribusi inferensi dan generatif untuk menghasilkan distribusi variasional rekursif dengan mekanisme perhatian seperti pada deep recurrent Attention Writer (DRAW). Dalam pembelajaran manipulasi teoretis permainan, mekanisme perhatian seperti itu menciptakan strategi respons terbaik bagi musuh sebagai keputusan operasional acak, sementara pengklasifikasi peka biaya mempelajari representasi untuk distribusi multimodal, multiview, dan multitask.

Zhou dkk. mengusulkan autoencoder mendalam untuk membedakan antara data yang direkonstruksi berkualitas tinggi dan outlier. Ia dapat menemukan anomali acak serta kerusakan terstruktur dengan algoritma deteksi anomali tanpa pengawasan. Ini merupakan perpanjangan dari autoencoder denoising dan autoencoder korentropi maksimum yang biaya rekonstruksinya adalah entropi tahan kebisingan. Dengan mendefinisikan proyeksi non-linier ke lapisan tersembunyi berdimensi rendah, autoencoder kuat yang diusulkan adalah versi non-linier dari analisis komponen utama yang kuat. Ini menghasilkan representasi data non-linier yang sesuai untuk menghasilkan tingkat kesalahan rekonstruksi yang lebih rendah pada distribusi input yang rumit. Metode pengganda arah bolak-balik adalah algoritma optimasi yang digunakan untuk melatih autoencoder. Autoencoder yang tangguh tersebut dapat digunakan untuk mendeteksi serangan siber pada data jaringan.

Lin, memberikan gambaran umum tentang metode agregasi peringkat. Mereka bertindak sebagai metode pencarian stokastik untuk menggabungkan kriteria optimasi yang berbeda dalam distribusi stasioner. Ukuran jarak digunakan untuk mengumpulkan daftar peringkat. Urutan elemen dalam daftar optimal ditentukan dalam matriks probabilitas yang diparameterisasi dengan kriteria divergensi informasi Monte Carlo lintas entropi antara pelatihan dan distribusi manipulasi. Gregor dkk. memperkenalkan jaringan autoregresif

mendalam untuk mempelajari hierarki representasi terdistribusi dari data. Algoritme estimasi parameter berdasarkan panjang deskripsi minimum (MDL) memaksimalkan batas bawah variasi pada estimasi kemungkinan log pada data pelatihan. Representasi pembuat enkode memainkan peran distribusi variasional yang ringkas dan tidak berlebihan dari sudut pandang teori informasi. Struktur autoregresif pada variabel laten menangkap ketergantungan antara unit aktivasi pada lapisan yang sama.

Dalam literatur statistik multivariat yang mempelajari sistem distribusi probabilitas, kopula memodelkan distribusi multivariat. Copula membangun distribusi gabungan dengan struktur ketergantungan yang berbeda seperti ketergantungan ekor yang dimodelkan sebagai distribusi marjinal. Kopula anggur memungkinkan estimasi kepadatan sewenang-wenang. Ukuran teori informasi berbasis entropi seperti informasi timbal balik dapat dibandingkan dengan kopula untuk mengukur ketergantungan multivariat antara fitur-fitur yang terlibat dalam regresi berganda terhadap distribusi data gabungan. Tagasovska dkk. memperkenalkan autoencoder vine copula untuk memperkirakan distribusi multivariat dari data yang dikodekan.

Model generatif menggabungkan perkiraan distribusi dengan decoder. Sebagai model generatif implisit, vine copula tidak menerapkan banyak batasan seperti metode variasional pada pelatihan di ruang laten. Mereka tidak membuat asumsi distribusi yang eksplisit pada decoder proses generatif data. Mereka bertindak sebagai alat yang fleksibel untuk membangun fitur dalam distribusi multivariat berdimensi tinggi. Data baru dapat dibuat dengan mendekode sampel acak yang dihasilkan oleh vine copula. Pemilihan keluarga kopula memungkinkan fleksibilitas dalam pemodelan dan eksplorasi nilai parameter dalam desain autoencoder menggunakan mekanisme pelatihan manipulasi.

Wieczorek dkk. menerapkan transformasi kopula di ruang laten autoencoder untuk membuat representasi fitur yang jarang. Representasi data seperti itu mengekstrak fitur-fitur yang ringkas, jarang, dan dapat ditafsirkan dalam pembelajaran mesin. Fitur renggang digabungkan dengan prinsip kemacetan informasi mendalam dalam inferensi variasional untuk mendapatkan batasan teori informasi pada efisiensi jaringan pembelajaran mendalam. Batas bawah variasional diturunkan dari masalah optimasi kemacetan informasi yang dirumuskan sebagai informasi timbal balik antara distribusi data adversarial dan data pelatihan. Ini melibatkan istilah entropi pada fitur diskrit, istilah entropi diferensial pada fitur kontinu, dan istilah entropi kopula marginal pada fitur laten. Augmentasi kopula dari autoencoder variasional diusulkan untuk memberikan ketahanan terhadap serangan musuh karena pengaruh positif pada tingkat konvergensi autoencoder.

Hua dkk. mengembangkan penduga dan filter peringkat rendah untuk komputasi subruang. Metode daya bolak-balik (AP) yang diusulkan untuk menghitung peringkat tereduksi secara komputasi lebih efisien dibandingkan metode yang ada dalam literatur. Pengurangan peringkat seperti itu terbukti dapat diterapkan dalam sistem multivariat dengan sejumlah besar sumber dan penerima di mana struktur internal dan interferensi sinyal multipath direpresentasikan dengan matriks saluran dengan peringkat yang dikurangi. Sistem seperti itu secara implisit perlu mengurangi kompleksitas model untuk mengimbangi beban

komputasi. Oleh karena itu, estimasi dan pemfilteran peringkat yang dikurangi berguna dalam berbagai aplikasi pemrosesan sinyal yang memerlukan reduksi data/model, ketahanan terhadap gangguan yang merugikan dan kesalahan pemodelan, serta efisiensi komputasi yang tinggi.

Di sini pembelajaran mendalam manipulasi dalam estimasi peringkat tereduksi dapat dibandingkan dengan metode yang lebih konvensional sebagai prosedur pembelajaran representasi yang efisien secara komputasi untuk estimasi peringkat seperti teknik dekomposisi nilai eigen (EVD), dekomposisi nilai tunggal (SVD), dan dekomposisi subruang (SSD). Lebih lanjut, metode AP yang diusulkan untuk optimasi merupakan alternatif dari metode pencarian gradien yang digunakan dalam optimasi pembelajaran adversarial. Ini memperluas pendekatan jarak minimum kuadratik berulang (IQMD) untuk mengoptimalkan fungsi kerugian dalam pembelajaran mendalam manipulasi.

Luedtke dkk. membangun meta-learning (AMC) Monte Carlo yang bermusuhan untuk prosedur statistik intensif komputasi dalam pendekatan frequentist dan Bayesian yang masing-masing memerlukan pengoptimalan penaksir kemungkinan maksimum dan pengambilan sampel dari distribusi probabilitas yang sulit dilakukan. Di sini masalah statistik dirumuskan sebagai permainan dua pemain di mana Alam secara bermusuhan memilih distribusi bagi ahli statistik untuk menjawab pertanyaan ilmiah dengan menggunakan data yang diambil dari distribusi ini. Solusi optimal ditemukan melalui strategi pemain yang diparameterisasi oleh jaringan saraf dalam. Kasus terburuk untuk kompleksitas pengambilan sampel dalam mekanisme penghasil data ditemukan ketika kinerja prosedur statistik paling tidak diinginkan.

Masalah kompleksitas pengambilan sampel tersebut diselesaikan dengan prosedur optimasi minimax yang secara statistik setara dengan prosedur Bayes yang diturunkan dari prior yang paling tidak menguntungkan pada kuantitas yang diminati yang paling sulit dihitung. Untuk menetapkan tingkat konvergensi dan jaminan kinerja pada prior yang paling tidak menguntungkan tersebut, metode otomasi dirancang pada kelas prior terbatas yang telah ditentukan sebelumnya pada serangkaian distribusi terbatas dalam masalah keputusan statistik. AMC yang diusulkan dapat digunakan untuk menggabungkan strategi manipulasi dalam penyesuaian, pemilihan, dan optimalisasi prosedur pembelajaran yang diawasi.

Mekanisme penghasil data yang bermusuhan dapat dibangun dari model statistik. Kerangka statistik untuk pembelajaran manipulasi diilustrasikan dalam tiga kelas masalah statistik: estimasi titik, prediksi, dan konstruksi wilayah kepercayaan. Risiko minimax dioptimalkan pada kelas prosedur pembelajaran yang memungkinkan. Secara umum, optimasi seperti itu sulit dilakukan pada waktu polinomial non-deterministik (NP). Berbagai strategi manipulasi dihipotesiskan untuk mengatasi kompleksitas komputasi dalam prosedur pembelajaran dengan optimasi numerik.

Salah satu strategi tersebut secara berulang meningkatkan risiko maksimal dari prosedur statistik. Algoritma minimax bersarang untuk menyusun prosedur minmax secara numerik adalah strategi lain. Strategi lain menggunakan algoritma bergantian untuk menghibridisasi algoritma minimax dan maximin yang bersarang. Oleh karena itu prosedur

statistik yang optimal untuk penambangan data dapat dibangun dengan pembelajaran mendalam yang bermusuhan terutama ketika prosedur statistik yang ada cenderung gagal.

Romano dkk. menganalisis stabilitas mesin klasifikasi pembelajaran mendalam seperti CNN. Mereka menemukan hubungan antara stabilitas klasifikasi terhadap kebisingan dan struktur yang mendasari sinyal. Tautan seperti itu dikuantifikasi dalam bentuk pembelajaran kamus atas representasi data yang jarang. Dengan demikian, bidang penelitian pembelajaran representasi renggang dan pembelajaran kamus dapat digunakan untuk menganalisis sensitivitas regresi dan ketahanan pengklasifikasi terhadap gangguan manipulasi. Kekokohan yang terikat pada energi kebisingan ditemukan sebagai fungsi dari ketersebaran sinyal dan karakteristiknya yang dinyatakan sebagai bobot representasi kamus. Jaringan parseval kemudian ditemukan sebagai regularisasi empiris untuk meningkatkan stabilitas klasifikasi. Solusi renggang dan kamus/filter yang tidak koheren pada sinyal yang masuk diusulkan sebagai solusi untuk membangun jaringan saraf yang kuat pada gangguan yang merugikan.

Guo dkk. mengungkapkan hubungan antara ketersebaran pengklasifikasi mendalam dan ketahanan manipulasi nya. Ketersebaran yang lebih tinggi menunjukkan ketahanan yang lebih baik dalam jaringan neural dalam non-linier. Pengklasifikasi renggang tidak hanya efisien secara komputasi tetapi juga menarik secara teoritis. Mereka dapat digunakan dalam desain mekanisme pertahanan dalam pembelajaran mendalam manipulasi seperti pelatihan manipulasi, penyulingan pengetahuan, pendeteksian dan penolakan, penyembunyian gradien, dan pengacakan. Di sini inefisiensi menyebabkan redundansi dalam desain pengklasifikasi mendalam dengan pemangkasan jaringan dan regularisasi tensor bobot.

Kreutz-Delgado dkk. mengembangkan pembelajaran berbasis data dari kamus khusus domain untuk kemungkinan maksimum dan estimasi a posteriori maksimum. Sebagai generalisasi kuantisasi vektor, elemen kamus ditafsirkan sebagai konsep, fitur, atau kata-kata yang mewakili peristiwa yang ditemui dan sinyal yang dihasilkan dalam lingkungan manipulasi. Evaluasi eksperimental menunjukkan bahwa algoritma pembelajaran representasi yang diusulkan berdasarkan pemecah sistem yang belum ditentukan berkinerja lebih baik daripada metode analisis komponen independen (ICA). Selain itu, kamus ini menghasilkan kompresi yang lebih tinggi (bit per piksel lebih sedikit) dan akurasi yang lebih tinggi (kesalahan kuadrat rata-rata lebih rendah).

Kamus yang bermakna bagi lingkungan diperoleh secara fisik atau biologis dengan memaksimalkan informasi timbal balik antara kumpulan vektor ini dan sinyal yang dihasilkan oleh lingkungan. Sekumpulan vektor atau kamus independen linier dengan rentang minimal mewakili sinyal terukur yang diinginkan dengan pengurangan kebisingan dan kompresi data. Mencocokkan sinyal sumber dengan kamus renggang juga dapat dipahami sebagai pemodelan entropi maksimum dari struktur statistiknya. Perkiraan kamus yang dihasilkan kemudian disebut sebagai perkiraan perkiraan kemungkinan maksimum dari sinyal sumber.

Literatur penyaringan adaptif untuk perkiraan kamus saat ini juga dapat digunakan untuk melacak sensitivitas pembelajaran terhadap gangguan manipulasi dengan koreksi berbasis data. Penggunaan kamus yang dipelajari juga dibandingkan dengan penggunaan kamus wavelet yang telah ditentukan sebelumnya untuk membuat ulang sinyal sensor yang

diamati dengan keterpisahan dan faktorisasi dalam distribusi data untuk pemodelan diskriminatif-generatif dalam pembelajaran mendalam adversarial teoretis permainan. Aplikasi ditemukan untuk fungsi kerugian multimodal yang dihasilkan, fungsi biaya multiview, dan fungsi tujuan multitask dalam pencitraan biomedis, suara seismik geofisika, dan pelacakan multitarget.

Zou dkk. analisis komponen utama renggang (SPCA) menggunakan laso untuk menghasilkan komponen utama yang dimodifikasi dengan pembebanan renggang. SPCA diformulasikan sebagai kerangka optimasi regresi dengan algoritma komputasi yang efisien pada data multivariat. Kriteria regresi yang mengidentifikasi variabel-variabel penting, bukan ambang batas sederhana pada varians yang dijelaskan, digunakan untuk memperoleh komponen-komponen utama utama. Tanpa batasan ketersebaran, metode ini direduksi menjadi PCA.

Sprechmann dkk. membuat kerangka pengelompokan dengan pembelajaran kamus dan pengkodean yang jarang. Titik-titik representatif untuk pengelompokan dimodelkan dalam bentuk distribusi data yang direpresentasikan dalam satu kamus untuk setiap cluster. Jadi seluruh konfigurasi pengelompokan dimodelkan sebagai gabungan subruang berdimensi rendah yang dipelajari dan titik datanya. Kamus yang dipelajari membuat kerangka pengelompokan tanpa pengawasan cocok untuk memproses kumpulan data besar dengan cara yang kuat. Algoritme optimasi berulang seperti EM dirancang untuk memisahkan cluster ke dalam kamus. Kamus juga digunakan dalam pengukuran kualitas representasi baru yang menggabungkan pengkodean renggang, pembelajaran kamus, dan pengelompokan spektral untuk pengelompokan keras dan lunak.

Pemrograman Matematis dalam Pembelajaran Permainan Strategis

Algoritma evolusioner (EA) telah digunakan dalam optimasi stokastik untuk menghasilkan model penambangan data berbasis aturan dengan interaksi atribut. Algoritma pencarian dan optimasi stokastik berbasis EA adalah Evolutionary Programming (EP), Evolutionary Strategy (ES), Genetic Algorithm (GA), Differential Evolution (DE), Estimation of Distribution Algorithm (EDA) dan Algoritma Swarm Intelligence (SI).

Dalam algoritme manipulasi kami, algoritme pencarian dan pengoptimalan dapat berupa algoritme genetika atau algoritme simulasi anil. Sampel data adversarial dihasilkan oleh operator pencarian seleksi, crossover, mutasi pada algoritma genetika dan operator pencarian anil pada algoritma simulasi anil. Dengan menggunakan algoritma pendakian bukit probabilistik melalui rantai Markov dalam model multivariat, operator pencarian saat ini dapat diperluas untuk menentukan distribusi probabilistik eksplisit yang melakukan pencarian lingkungan yang kompleks untuk kandidat solusi.

Harada dkk. menganalisis kemajuan dalam algoritma genetika paralel (PGA). PGA dapat digunakan sebagai algoritme pengoptimalan ketika fungsi sasaran target tidak dapat diturunkan, tidak kontinu, dan tidak terdefinisi dengan baik, serta tidak memiliki ekspresi analitis apa pun. Mereka dapat menggabungkan ruang pencarian berdimensi tinggi, operator yang disesuaikan untuk aplikasi, kumpulan data kompleks dalam algoritma pencarian, dan batasan non-linier pada tujuan optimasi. Di sini mereka dapat memperoleh keuntungan dari

paralelisasi dan platform pemrosesan terdistribusi seperti multiprosesor, GPU, FPGA, cluster, grid, dan cloud dengan menghemat pencarian dan pengoptimalan, menghemat pemanggilan fungsi dan komputasi numerik.

API pengoptimalan untuk mengimplementasikan PGA dapat dikategorikan menjadi komputasi paralel, komputasi terdistribusi, MPI, dan CUDA. Implementasi tersebut berjalan pada uniprosesor, komputer paralel, dan jaringan stasiun kerja. Mereka memungkinkan pengembangan algoritma optimasi canggih untuk algoritma tujuan tunggal, multi-tujuan, dan paralel dalam arsitektur berorientasi objek. PGA memecahkan masalah dalam aplikasi dunia nyata seperti penambangan data, pencarian jalur, lalu lintas jalan raya, perencanaan penggunaan lahan, ilmu nano, elektronik, struktur bangunan, dan sistem tenaga. PGA berguna untuk pemilihan fitur, optimasi hyperparameter, dan rekayasa fitur dalam penambangan data. Mereka mengarah pada penerapan analisis data besar, pembelajaran mendalam, kecerdasan komputasi, dan asal data dengan pembelajaran mesin yang berlawanan.

Area aktif untuk penelitian dalam PGA adalah skalabilitas terhadap kumpulan data berdimensi tinggi, ketahanan hasil optimasi terhadap perubahan parameter algoritma karena ketidakpastian dalam data dan lingkungan pembelajaran yang dinamis, evaluasi fungsi multi-tujuan untuk membangun secara efisien beragam dan tinggi. -solusi berkualitas dalam pengambilan keputusan multikriteria, analisis trade-off algoritmik seperti kegunaan/efisiensi dalam merancang pencarian, algoritma paralel dan metrik pembelajaran dalam solusi big data, pemrosesan data dan analisis algoritmik dalam PGA pada perangkat dan layanan komputasi fog/edge, PGA untuk komputasi berkinerja tinggi yang menggabungkan algoritme eksak/perkiraan untuk kebijakan komunikasi sinkron/asinkron, dan arsitektur layanan mikro yang membangun solusi kompleks dengan PGA yang bertindak sebagai layanan web yang menyediakan optimasi tervalidasi.

Algoritme evolusi paling populer untuk pengoptimalan pembelajaran mesin adalah Stochastic hill-climbing, simulasi anil, dan algoritma genetika. Goldberg membahas penggunaan algoritma genetika dalam optimasi stokastik pembelajaran mesin. Mekanisme evolusi disimulasikan dalam komputer dengan populasi data yang berisi karakteristik solusi yang dikembangkan dari generasi ke generasi dari populasi tersebut untuk melatih model pembelajaran mesin dalam lingkungan dengan fungsi objektif untuk mengoptimalkan populasi. Ini adalah algoritma optimasi berulang di mana setiap solusi individual ditandai dengan nilai fungsi kebugaran. Solusinya akan tercapai jika populasi dan tujuan pembelajarannya didefinisikan dengan baik.

Di sini algoritma genetika bekerja pada populasi dengan banyak kemungkinan solusi secara bersamaan. Mereka perlu menghitung nilai fungsi kebugaran tanpa memerlukan informasi tambahan seperti turunan dari fungsi tujuan. Mereka kemudian menggunakan aturan pembaruan probabilistik untuk mengembangkan pengacakan menjadi solusi kandidat. Di sini pembelajaran mendalam dapat digunakan untuk merepresentasikan data pelatihan sebagai populasi solusi untuk algoritma genetika. Perluasan algoritma genetika ke optimasi multi-tujuan menghasilkan solusi Pareto-optimal. Konsep optimasi dalam algoritma genetika

dapat diperluas untuk memilih tidak hanya parameter pemodelan tetapi juga fungsi kebugaran dan teknik optimasi sebagai bagian dari masalah pembelajaran mesin adversarial.

Michalewicz mensurvei teknik pemrograman evolusioner untuk menggabungkan pengetahuan khusus masalah sebagai operator khusus dalam algoritma genetika. Mereka mengarah pada program evolusi yang merupakan algoritma probabilistik yang memperluas prinsip-prinsip algoritma genetika. Operator khusus dapat digunakan untuk optimasi numerik, penyetelan model, pencarian terbatas, pembelajaran strategi, dan optimasi multimodal dalam pembelajaran manipulasi teoritis permainan. Di sini, jaringan pembelajaran mendalam dapat melampaui pengkodean biner populasi untuk merepresentasikan fitur pembelajaran mesin dalam operator komputasi fuzzy, numerik, untuk program evolusi.

Para pemain yang terkait dengan strategi tertentu dalam pemodelan teoritis permainan dapat direpresentasikan sebagai populasi dalam program evolusi. Fungsi imbalan adversarial kemudian dapat bertindak sebagai fungsi kesesuaian yang mengevaluasi solusi individual yang akan dipilih untuk generasi berikutnya. Strategi yang lebih baik dapat dibangun dengan mengawinkan pemain lintas generasi. Representasi strategi dapat diacak dengan operator genetik. Minimalkan penyesalan seorang pemain ditentukan oleh rata-rata pembayaran yang diterimanya atas semua permainan yang dimainkannya. Dengan cara ini program evolusi dapat digunakan untuk menyelesaikan permainan multi-label multi-label dalam pembelajaran adversarial yang diawasi dengan optimalisasi simultan berbagai tujuan dalam masalah pengambilan keputusan di dunia nyata.

Pembelajaran empiris simbolik adalah bidang penelitian dalam pemrograman evolusioner yang dapat menyebabkan aturan klasifikasi untuk pembelajaran yang diawasi. Berbeda dengan sistem pengklasifikasi simbolik yang mempertahankan pengetahuan eksplisit dalam bahasa deskriptif tingkat tinggi, model statistik mewakili pengetahuan sebagai serangkaian contoh dan statistik yang terkait dengannya, dan model koneksionis mewakili pengetahuan di antara bobot yang ditetapkan pada koneksi jaringan saraf. Pembelajaran empiris simbolik yang diterapkan pada sistem pengklasifikasi harus mendefinisikan sistem berbasis aturan seperti sistem detektor-efektor untuk menyandikan-mendekode data pelatihan menjadi representasi solusi genetik, sistem pesan pada masukan ke algoritma genetika, sistem aturan yang menghasilkan populasi pengklasifikasi, sistem kredit untuk mengembangkan solusi lintas generasi, dan prosedur genetik untuk menghasilkan populasi untuk berbagai sistem berbasis aturan.

Di sini program evolusi dapat digunakan untuk memodelkan perilaku skenario serangan teoretis permainan dalam pembelajaran adversarial yang diawasi. Representasi fitur khusus masalah dan operator khusus untuk program evolusi dapat menerapkan algoritma evolusioner dalam mesin keadaan terbatas untuk optimasi numerik, pembelajaran mesin, permainan berulang, kontrol optimal, pemrosesan sinyal, pemodelan kognitif, desain teknik, integrasi sistem, dan robotika. Osilasi strategis adalah pendekatan optimasi terbatas yang dapat diterapkan pada masalah optimasi kombinatorial dan non-linier yang diselesaikan dengan program evolusi. Ini melampirkan konteks kelayakan/ketidaklayakan pada desain

pencarian lingkungan yang sensitif terhadap biaya dan optimalisasi stokastik dalam program evolusi.

Konfigurasi sistem berbasis aturan untuk memilih wilayah yang akan dilalui dan arah traversal ditentukan oleh kemampuan untuk mendekati dan melintasi batas kelayakan dari arah yang berbeda. Menelusuri kembali lintasan sebelumnya dihindari oleh mekanisme memori dan probabilitas. Proses konstruktif untuk mencapai batas kelayakan dibarengi dengan proses destruktif dalam membongkar strukturnya sehingga mengakibatkan fluktuasi strategis di sekitar batas tersebut. Osilasi strategis seperti itu dapat digunakan untuk memandu peningkatan fungsi imbalan yang merugikan di sekitar batas pengklasifikasi dengan prosedur pencarian yang menyelidiki kedalaman wilayah terkait. Batasan masalah pada pencarian tersebut dapat membatasi dan menghukum pencarian dengan batasan yang ditetapkan pada fungsi bernilai vektor.

Pengorbanan antara tingkat pelanggaran yang berbeda terhadap batasan komponen dapat diperbolehkan sesuai dengan skor kepentingan fiturnya. Masalah seperti ini disebut masalah kepuasan kendala dalam program evolusi dan sebanding dengan teknik pemrograman kendala dalam optimasi matematis. Oleh karena itu, pembelajaran mendalam adversarial teoretis permainan dapat memperoleh manfaat dari teknik evolusi untuk optimalisasi fungsi dengan sistem yang dapat beradaptasi sendiri yang menggabungkan parameter kontrol ke dalam vektor solusi, sistem ko-evolusi di mana proses evolusi terhubung antar populasi, struktur poliploid yang menggabungkan memori lingkungan non-stasioner ke dalam solusi individual , dan model pemrograman paralel besar-besaran yang menyematkan komputasi evolusioner.

McCune dkk. menyajikan survei model pemrograman vertex-centric dalam kerangka pemrosesan terdistribusi untuk jaringan yang kompleks. Ini terdiri dari komponen yang saling bergantung untuk menghitung algoritma grafik berulang dalam skala besar. Dengan demikian kita dapat mengevaluasi sensitivitas fungsi kerugian yang merugikan sehubungan dengan penemuan struktur konektivitas, representasi, visualisasi, dan evaluasi pemodelan teoritis permainan pada jaringan yang kompleks. Dalam konteks pembelajaran mesin manipulasi atas dinamika penambangan pola grafik, kita dapat mengeksplorasi konstruksi pemrograman fungsional yang cocok untuk pemrosesan terdistribusi seperti MapReduce dan paralel sinkron massal.

Pilihan model pemrograman adalah antara paralelisme data, paralelisme tugas, dan paralelisme grafik. Haller dkk. membahas tantangan penerapan pembelajaran mesin paralel dan terdistribusi dengan abstraksi pemrograman fungsional. Detail implementasi untuk analisis data terdistribusi harus mempertimbangkan asumsi pembelajaran mesin yang tersirat dalam model data, model memori, model pemrograman, model komunikasi, model eksekusi, dan model komputasi algoritma paralel dan serial. Metode pembelajaran fitur yang relevan meliputi sampel, pohon, cluster, wavelet, kernel, splines, jaring, filter, wrapper dan faktor dalam rangkaian data, urutan, grafik, dan jaringan. Pemodelan teoritis permainan perlu mempelajari substruktur padat, kelas langka, dan pola padat pada kumpulan data

transaksional, sekuensial, dan grafik di mana proses acak yang menghasilkan data pelatihan mungkin tidak sama dengan yang mengatur data pengujian.

Miller dkk. membahas model pemrograman paralel yang dibuat khusus untuk pembelajaran mesin yang diimplementasikan dalam bahasa pemrograman Scala. Mereka harus mendukung pemrosesan grafik terdistribusi, menyediakan operasi massal paralel pada koleksi umum, dan membuat bahasa khusus domain paralel untuk pembelajaran mesin pada platform perangkat keras yang heterogen. Fitur bahasa Scala untuk merancang dan mendistribusikan sistem waktu berjalan paralel untuk pembelajaran mesin juga dibahas. Untuk pemodelan teoretis permainan, kita harus merancang mekanisme pembelajaran tanpa pengawasan dengan model penambangan motif seperti pengelompokan biclustering dan evolusioner, pengelompokan bertingkat, dan pengelompokan berbasis model.

Untuk membuat model pembelajaran antagonis terawasi dengan motif seperti itu, kita dapat fokus pada metode kompresi dan metode optimasi dalam pembelajaran kernel dan pembelajaran mendalam. Teori yang relevan dalam penambangan data adalah pengelompokan bertingkat, partisi grafik bertingkat, deteksi kuasi-klik, dan penemuan subgraf padat. Struktur pengindeksan data untuk data dinamis juga akan mengurangi biaya komunikasi dan meningkatkan penyeimbangan beban dalam sistem memori terdistribusi tersebut.

Mohammed dkk. membangun sistem klasifikasi fuzzy dengan simulasi anil. Pengetahuan yang ditemukan berupa aturan prediksi if-then dari representasi pengetahuan simbolik. Mereka dapat dievaluasi berdasarkan beberapa kriteria signifikansi statistik seperti tingkat kepercayaan dalam prediksi, tingkat akurasi klasifikasi pada instance kelas yang tidak diketahui, dan kemampuan interpretasi dari metode penalaran perkiraan sistem fuzzy. Setiap fungsi keanggotaan yang disesuaikan dapat dikembangkan dalam sistem klasifikasi fuzzy untuk masalah klasifikasi pola tertentu. Simulasi anil melakukan pencarian global terhadap masalah klasifikasi untuk keluar dari optimum lokal. Algoritma genetika rata-rata membutuhkan waktu polinomial.

Beyer dkk. melakukan analisis kompleksitas algoritma evolusioner dalam ruang pencarian kontinu dan diskrit. Pendekatan teoritis terhadap desain dapat membantu kita memahami dan mengajarkan algoritma evolusioner sebagai metode optimasi probabilistik dalam kecerdasan komputasi. Analisis pertama berkisar pada ukuran kinerja yang disebut kecepatan kemajuan. Ini adalah jarak rata-rata dalam ruang pencarian yang ditempuh dalam arah yang berguna per evaluasi fungsi. Adaptasi diri dan kemungkinan keberhasilan yang terkait dengan kekuatan mutasi menganggap fungsi tujuan sebagai kotak hitam untuk optimasi. Dalam algoritma blackbox, semua penghitungannya gratis, dan hanya pengambilan sampelnya yang dikenakan biaya. Terakhir, kebaikan statistik dari solusi yang ditemukan oleh algoritme evolusioner bergantung pada pengetahuan atau ketidaktahuan tentang karakteristik masalah yang direpresentasikan sebagai fitur pembelajaran mesin untuk pengoptimalan.

Untuk memperhitungkan masalah desain tersebut, analisis kompleksitas memprediksi perilaku algoritma evolusioner setelah sistem dinamis. Konsep tatanan konvergensi dalam teori optimasi memberikan batasan pada gangguan kebugaran dan dinamika evolusi. Sebagai algoritma acak, kita dapat mengaitkan probabilitas keberhasilan mencapai optimal dengan algoritma evolusioner. Mengingat distribusi probabilitas masukan, pengacakan memiliki batas bawah untuk perkiraan waktu optimalisasi kasus terburuk. Dalam praktiknya, mereka mengembangkan solusi perkiraan di bawah batasan perangkat keras. Dalam hal ini algoritma evolusioner dapat dirancang sebagai teknik perbaikan.

Xue dkk. mensurvei komputasi evolusioner mutakhir sebagai teknik pencarian global untuk pemilihan fitur dalam ruang pencarian yang besar. Pemilihan fitur seperti itu diterapkan dalam beberapa tugas pembelajaran mesin seperti klasifikasi, pengelompokan, regresi, dan prediksi. Algoritma genetika, optimasi kawanan partikel, dan optimasi koloni semut adalah metode komputasi evolusioner yang paling populer dalam pemilihan fitur. Mereka dapat diintegrasikan dan dimasukkan ke dalam pembelajaran pengklasifikasi sebagai pendekatan tertanam untuk pemilihan fitur. Kemudian pemrograman genetika bertindak sebagai teknik optimasi untuk pembelajaran mesin. Sistem pengklasifikasi pembelajaran dapat memanfaatkan pemilihan fitur yang tertanam. Di sini interaksi fitur dengan konsep target dievaluasi dengan subset fitur optimal yang dievaluasi dengan komputasi evolusioner.

Algoritme evolusioner berperan dalam pencarian subset fitur serta kriteria evaluasinya yang tujuannya adalah untuk memaksimalkan akurasi klasifikasi sekaligus meminimalkan jumlah fitur. Oleh karena itu pemilihan fitur dengan metode evolusioner dapat diperlakukan sebagai masalah multi-tujuan yang harus menemukan serangkaian solusi trade-off yang tidak didominasi. Mereka tidak perlu membuat asumsi tentang ruang pencarian seperti apakah ruang tersebut dapat dipisahkan dan dibedakan secara linier atau non-linier. Mekanisme berbasis populasi mereka dapat menghasilkan banyak solusi dalam satu proses yang dapat diparalelkan. Namun ada kebutuhan untuk meningkatkan stabilitas algoritma evolusioner yang cenderung memilih fitur berbeda dari proses yang berbeda.

Masalah desain ini juga meningkatkan kompleksitas komputasi algoritma evolusioner pada tugas dunia nyata dengan banyak fitur. Ukuran evaluasi kinerja komputasi evolusioner dalam pembelajaran mesin bersumber dari teori informasi, ukuran korelasi, ukuran jarak, teori himpunan fuzzy, dan teori himpunan kasar. Mereka mengarah pada aplikasi dalam pemrosesan gambar dan sinyal, pengenalan wajah, pengenalan tindakan manusia, pengenalan pembicara, pengenalan angka tulisan tangan, identifikasi pribadi, deteksi biomarker, diagnosis penyakit, deteksi spam email, keamanan jaringan, pembelajaran bahasa, dan optimalisasi sistem tenaga.

Suman dkk. meninjau algoritma optimasi berdasarkan simulasi anil untuk masalah optimasi tujuan tunggal dan multi. Perhitungan probabilitas untuk membangun jadwal anil dibahas di berbagai algoritma. Jadwal anil dapat digunakan untuk mendapatkan serangkaian solusi Pareto untuk masalah optimasi multi-tujuan. Sebuah studi tentang hasil komputasi dan lingkungan kinerja simulasi anil dapat menyarankan perbaikan pada jadwal anil. Simulated annealing membutuhkan waktu lebih sedikit dibandingkan algoritma genetika karena

algoritma ini menemukan solusi optimal melalui iterasi poin demi poin dibandingkan pencarian pada populasi individu. Ini dapat dianggap sebagai pendekatan heuristik acak untuk masalah optimasi kombinatorial seperti masalah travelling salesman. Ini dapat secara efisien mengakomodasi tujuan desain yang beragam dan saling bertentangan dalam masalah optimasi multi-tujuan seperti tata letak sirkuit terpadu.

Metode berbasis gradien dan Hessian tidak efektif dalam aplikasi pemrosesan sinyal yang melibatkan masalah optimasi dengan fungsi kerugian multimodal dan tidak mulus. Anil simulasi adaptif adalah alat optimasi dalam masalah optimasi non-linier. Hal ini dapat diterapkan tidak hanya untuk masalah optimasi tetapi juga dalam klasifikasi objek dan pengenalan pola dimana metrik jarak dapat menjadi fungsi tujuan. Simulasi anil dapat dikombinasikan dengan algoritme genetika untuk memberikan solusi efisien yang mengeksplorasi lingkungan simulasi anil dalam masalah multikriteria di mana algoritme genetika menyesuaikan parameter yang disetel untuk setiap tujuan di setiap iterasi. Algoritme genetika tersebut dapat memodelkan populasi sampel dari solusi yang berinteraksi sementara simulasi anil menerima solusi yang layak dengan beberapa probabilitas yang ditentukan oleh jadwal anil.

Deb membahas tantangan dalam optimasi multi-tujuan terutama ketika tujuan-tujuan tersebut bertentangan satu sama lain. Kemudian mereka memunculkan solusi optimal tradeoff dengan atau tanpa kendala optimasi yang disebut solusi Pareto-optimal. Optimasi multi-tujuan evolusioner adalah bidang penelitian yang mempelajari masalah-masalah tersebut. Berbeda dengan metode berbasis gradien, optimasi multi-tujuan evolusioner tidak memerlukan informasi turunan apa pun untuk menemukan solusi optimal. Ini dapat memecahkan masalah multimodal dan menormalkan variabel keputusan dengan populasi yang terus berkembang dengan memanfaatkan nilai minimum dan maksimum fungsi tujuan dan kendala. Ini dapat menggabungkan operator stokastik dan deterministik yang cenderung menyatu ke solusi yang diinginkan dengan probabilitas tinggi.

Operator tersebut meliputi seleksi, persilangan, mutasi, dan pelestarian elit. Penggunaan populasi data dalam mekanisme pencarian optimasi evolusioner secara implisit dapat menerima pemrograman paralel yang memalukan di berbagai wilayah ruang pencarian. Hal ini dapat memecahkan permasalahan optimasi dunia nyata yang melibatkan tujuan yang tidak dapat dibedakan dan batasan yang tidak kontinu, solusi non-linier, keleluasaan, skala, pengacakan dalam perhitungan, dan ketidakpastian dalam keputusan. Konsep matematika yang disebut pengurutan parsial mendefinisikan solusi optimal Pareto yang tidak mendominasi dalam optimasi multi-tujuan evolusioner. Kriteria konvergensi optimasi multi-tujuan evolusioner dapat dikombinasikan dengan teknik optimasi matematis untuk menghasilkan pengoptimal dinamis. Algoritme optimasi multiobjektif evolusioner tersebut dapat dijelaskan sehubungan dengan aplikasi seperti desain lintasan pesawat ruang angkasa.

Mereka dievaluasi dengan ukuran kinerja pada bagian depan Pareto-optimal seperti rasio kesalahan, jarak dari kumpulan referensi, hypervolume, cakupan, R-metrik, dll. Algoritme optimasi multi-tujuan evolusioner dapat menangani stokastik dalam parameter masalah, variabel keputusan, fitur dimensi, dan sifat konvergensi dengan penilaian

probabilistik dari nilai fungsi tujuan dan kendala yang menemukan solusi yang tidak tepat dalam lingkungan yang tidak pasti. Prosedur seperti ini disebut metode pemrograman stokastik yang mengarah ke batas ketahanan (robustness frontier) dalam solusi optimal. Hal ini secara praktis diselesaikan dengan formulasi optimasi dua tingkat di banyak bidang sains dan teknik.

Kelley melakukan analisis matematis terhadap kondisi perlu dan cukup dalam optimasi berulang. Algoritme pengoptimalan untuk tujuan yang berisik dan batasan yang dibatasi dirangkum. Sra dkk. membahas peran metode pengoptimalan dalam pembelajaran mesin. Metode penurunan gradien stokastik dirangkum untuk optimasi skala besar cembung tidak mulus. Metode minimalisasi penyesalan diusulkan untuk memilih, mempelajari, dan menggabungkan fitur guna mengoptimalkan fungsi kerugian dalam pembelajaran mesin. Kebutuhan akan perkiraan optimasi dan analisis asimtotiknya diberikan untuk pembelajaran mesin skala besar. Akhirnya disajikan hubungan pembelajaran ketahanan dan kesalahan generalisasi serta perannya dalam optimasi kuat dengan pembelajaran manipulasi. Pengoptimalan online dan pengoptimalan bandit diusulkan sebagai metode untuk menangani gangguan manipulasi dan kebisingan label dalam pembelajaran yang diawasi.

Koziel dkk. mengulas bidang penelitian yang disebut optimasi komputasi. Model dan algoritma optimasi komputasi mencoba memanfaatkan sumber daya yang tersedia secara optimal untuk memaksimalkan keuntungan, keluaran, kinerja, dan efisiensi sekaligus meminimalkan biaya dan konsumsi energi. Algoritma pencarian adalah alat praktis untuk mencapai solusi optimal dalam optimasi komputasi. Mereka harus mengatasi ketidakpastian dalam sistem dunia nyata dengan desain yang kuat untuk fungsi tujuan dalam optimasi komputasi. Teknik optimasi cembung yang banyak digunakan dalam pembelajaran mesin adalah kasus khusus optimasi komputasi. Desain yang memuaskan untuk ketahanan harus menciptakan metodologi optimasi yang dapat dilakukan dengan sumber daya komputasi yang terbatas dan tujuan analitis yang sulit dilakukan. Metodologi optimasi tersebut terdiri dari komponen model, pengoptimal, dan simulator.

Model matematika atau numerik merupakan representasi dari permasalahan dunia nyata. Pengoptimal adalah algoritma yang menemukan solusi optimal. Dalam proses pencarian, pengoptimal menghasilkan dan mencari solusi baru dari solusi yang diketahui. Evaluator atau simulator adalah alat komputasi yang efisien dalam pemanfaatan waktu dan biaya komputasi secara keseluruhan. Ini biasanya terlibat dalam evaluasi nilai fungsi tujuan. Teorema tidak ada makan siang gratis untuk pembelajaran mesin dan pengoptimalan menyatakan bahwa tidak ada kemungkinan satu model universal, pengoptimal, dan simulator dapat diterapkan untuk semua variasi masalah pengoptimalan. Di sini algoritma optimasi dan perbaikannya dapat dikategorikan ke dalam metode berbasis turunan atau metode bebas turunan, metode berbasis lintasan atau berbasis populasi, metode deterministik atau stokastik, metode tanpa memori atau berbasis sejarah, dan metode lokal atau global.

Algoritme pengoptimalan bebas turunan dibandingkan dengan plot lintasan. More dkk. mengusulkan profil data sebagai alat pembandingan untuk menganalisis kinerja pemecah optimasi bebas turunan ketika anggaran komputasi untuk masalah pembandingan

terbatas. Mereka dapat dikombinasikan dengan kriteria konvergensi untuk mengevaluasi penurunan nilai fungsi tujuan dengan evaluasi fungsi yang mahal dalam masalah yang mulus, berisik, dan mulus sedikit demi sedikit. Profil kinerja mengevaluasi kinerja pemecah untuk berbagai tingkat akurasi diskriminatif. Biaya komputasi yang dominan disebabkan oleh jumlah evaluasi fungsi per iterasi. Profil kinerja dan profil data adalah fungsi distribusi kumulatif yang membandingkan pemecah yang berbeda. Tidak seperti profil kinerja, profil data mengekspresikan anggaran komputasi untuk mencapai pengurangan nilai fungsi tertentu dalam bentuk gradien simpleks untuk semua pemecah. Plot data melengkapi ukuran kinerja relatif plot kinerja dengan anggaran komputasi. Kyrola dkk. mengusulkan algoritma penurunan koordinat paralel untuk meminimalkan kerugian teratur L1 yang disebut Shotgun. Sebuah studi empiris tentang Shotgun dilakukan dalam masalah regresi logistik Lasso dan sparse. Ini terbukti menjadi metode pengoptimalan yang terukur.

Oliehoek dkk. menyajikan konsep solusi teori permainan untuk menjamin kemajuan dalam algoritma ko-evolusi. Algoritme ko-evolusi adalah pendekatan komputasi evolusioner yang mencari solusi optimal untuk masalah berbasis pengujian tanpa perlu menentukan fungsi kebugaran. Dalam pembelajaran manipulasi teoritis permainan, algoritma ko-evolusi dapat menganalisis konsep solusi yang diinginkan yang memaksimalkan utilitas yang diharapkan dalam pembelajaran dalam permainan, pembelajaran konsep, perkiraan dan klasifikasi fungsi, dan klasifikasi kepadatan menggunakan automata seluler. Pareto-co-evolution mengasosiasikan setiap pengujian dengan tujuan terpisah di mana himpunan solusi yang tidak didominasi adalah konsep solusi yang ditentukan. Dalam sistem multi-agen, keseimbangan Nash adalah konsep solusi yang ditentukan dalam permainan dua pemain. Ini merekomendasikan strategi campuran acak untuk setiap pemain yang berpartisipasi dalam permainan.

Memori Nash Paralel disajikan oleh Oliehoek dkk. untuk menganalisis konsep solusi dalam permainan asimetris. Ini merekomendasikan strategi respons terbaik dari proses keputusan Markov yang dapat diamati sebagian yang dibangun untuk permainan bentuk ekstensif yang terbatas. Representasi bentuk ekstensif menunjukkan representasi pohon dari interaksi teoretis permainan dalam suatu aplikasi. Algoritme ko-evolusi digunakan tidak hanya sebagai heuristik pencarian tetapi juga sebagai pengujian mekanisme memori di Parallel Nash Memory. Ia mampu mengidentifikasi strategi respons terbaik. Proses keputusan Markov kemudian diselesaikan dengan menggunakan teknik pemrograman dinamis seperti iterasi nilai. Kemudian prosedur Memori Nash Paralel direduksi untuk mengoordinasikan pendakian atau maksimalisasi bergantian.

Cai dkk. mengusulkan generalisasi multipemain dari game minimax zero-sum ke game polimatriks zero-sum. Permainan polimatriks ditentukan oleh graf yang simpulnya adalah para pemainnya dan sisi-sisinya adalah permainan dua pemain. Mengingat profil strategi untuk semua pemain, imbalan untuk setiap simpul adalah jumlah imbalan semua permainan dalam daftar kedekatannya. Imbalan untuk semua pemain berjumlah nol. Kesetimbangan teoretis permainan yang berbeda memberikan hasil yang berbeda kepada pemain dalam

permainan polimatriks jumlah nol. Algoritma pembelajaran tanpa penyesalan digunakan untuk mencari keseimbangan Nash dalam permainan polimatriks jumlah nol.

Bertsekas dkk. memberikan metode numerik untuk komputasi paralel dan terdistribusi dalam pemodelan teoritis permainan. Pemrograman Dinamis, proses keputusan Markov, dan algoritma optimasi stokastik dapat digunakan untuk merancang metode iteratif asinkron dalam pembelajaran adversarial teoritis permainan. Bisseling dkk. menjelaskan model paralel sinkron massal dan antarmuka penyampaian pesannya untuk pemrograman paralel yang memalukan dalam komputasi ilmiah untuk pembelajaran mendalam manipulasi teoretis permainan. Biaya optimasi komputasi pemrograman paralel dengan perkiraan tensor peringkat rendah dinyatakan sebagai biaya komputasi, biaya komunikasi, dan biaya sinkronisasi.

Penelitian algoritma komputasi untuk optimasi menciptakan metode numerik dengan karakterisasi terbaik dari fitur pembelajaran mesin dalam solusi dan kemudian mengurangi estimasi kesalahan pemodelan untuk kelas kompleksitas merancang solusi optimal. Jumlah pekerjaan komputasi yang terlibat dalam mewujudkan algoritma tergantung pada informasi dalam tipe data yang dipertimbangkan. Ini tidak hanya terdiri dari pekerjaan komputasi yang diperlukan untuk memperoleh informasi tentang data awal, tetapi juga jumlah pekerjaan yang diperlukan dalam memproses informasi dalam model pembelajaran mesin. Di sini pembelajaran manipulasi teoritis permainan menganalisis stabilitas dan ketahanan algoritma pembelajaran mesin dalam kaitannya dengan istilah utama estimasi kesalahan komputasi.

Algoritme genetika dalam pemodelan teoritis permainan dapat dikombinasikan dengan teknik optimasi seperti simulasi anil yang memanfaatkan metode pencarian garis dan wilayah kepercayaan untuk membuat manipulasi. Dalam studi tentang sifat konvergensi permainan optima teoritis, algoritma genetika juga dapat diganti dengan algoritma optimasi stokastik bebas turunan seperti pencarian pola, pencarian koordinat bertingkat, dan evolusi diferensial. Metode komputasi numerik dalam teori permainan evolusioner dan diferensial dapat menambah fungsi pembayaran manipulasi dengan persamaan keadaan diferensial parsial dari sistem dinamis untuk menghasilkan kontrol stokastik dalam interaksi teoritis permainan. Diferensial, perbedaan, dan persamaan aljabar Stackelberg Riccati dari permainan kesetimbangan teoretis dapat dimodelkan sebagai skenario serangan pembelajaran adversarial dengan musuh variasional dan generatif.

Pengoptimalan matematis dalam kondisi ketidakpastian dapat menjadi pandangan yang diambil sebagai mekanisme pertahanan untuk pembelajaran mesin yang dibuat tangguh dengan menyertakan algoritme pengoptimalan seperti pemrograman stokastik, pemrograman non-linier, pemrograman fuzzy, pengoptimalan tangguh adaptif, dan pengoptimalan tangguh berbasis data di dalamnya. prosedur pelatihan manipulasi. Peningkatan kapasitas pembelajaran, strategi pengacakan, dan fungsi pembayaran yang berpusat pada privasi dalam formulasi teoretis pengklasifikasi game akan memengaruhi pembobotan ulang regularisasi dan batasan keputusan terkait algoritme pembelajaran mesin yang disediakan sebagai layanan cloud.

Pendekatan Tingkat Rendah dalam Pembelajaran Permainan Strategis

Musuh dapat mengeksplorasi pemfilteran sinyal, deteksi, dan estimasi dalam tensor untuk mengekspresikan ketahanan, keadilan, penjelasan, dan transparansi pembelajaran mesin. Di sini representasi tensor dari distribusi data pelatihan di jaringan pembelajaran mendalam mengeksplorasi struktur dan konteks yang mendasari data dengan teori pembelajaran dan pengoptimalan berdasarkan sensitivitas aljabar tensor dari fungsi kerugian dalam pembelajaran mesin. Tensor dapat dipahami sebagai array multidimensi. Setiap arah dalam tensor disebut mode. Banyaknya fitur dalam suatu mode disebut dimensi. Peringkat tensor adalah jumlah total indeks kovarian suatu tensor. Rank adalah jumlah minimal mode dalam sebuah tensor. Peringkat tidak bergantung pada jumlah dimensi ruang fitur yang mendasari tensor. Pangkat suatu tensor disebut juga urutan atau derajat tensor. Dalam berbagai aplikasi, tensor didekomposisi menjadi tensor tingkat rendah menggunakan aljabar abstrak.

Dari perspektif pembelajaran mesin yang diawasi, aljabar tensor dapat didasarkan pada teori pembelajaran komputasi model pembelajaran mesin dan tugas penambangan data. Dalam pembelajaran adversarial teoretis permainan, penerapan dekomposisi tensor menarik dalam studi tentang tradeoff bias-varians dalam fungsi hasil adversarial untuk optimasi matematis. Kita dapat mencoba menjelaskan dekomposisi tensor dalam manipulasi untuk mempelajari efek bias algoritmik dalam pembelajaran mendalam.

Selanjutnya teori optimasi yang kuat dapat diusulkan untuk pembelajaran mendalam adversarial berbasis pengacakan. Teori pembelajaran mendalam seperti itu juga akan diterapkan dalam tugas penambangan data seperti deteksi kebaruan dan ekstraksi fitur. Di sini mesin faktorisasi adalah perkiraan tingkat rendah dari rekayasa fitur dalam tensor data renggang ketika sebagian besar elemen prediksinya tidak diketahui. Di sini komputasi granular berguna untuk membuat aturan fusi data pada representasi fitur data pelatihan. Hal ini dapat menyebabkan sistem neuro-fuzzy dan sistem multi-agen dalam penambangan data.

Kita dapat menyelidiki lebih lanjut transfer aturan fusi data yang signifikan secara statistik antara representasi data prediktif pada resolusi spasial dan distribusi data resolusi spektral dari manifold pelatihan. Struktur kompleks data temporal dalam keamanan siber juga dapat direpresentasikan sebagai grafik multidimensi dinamis untuk pembelajaran positif tanpa label. Grafik tersebut dapat diartikan sebagai jaringan kompleks dan tensor kompleks dalam penambangan data. Mereka memerlukan penggunaan pemrosesan data besar yang terdistribusi untuk penambangan grafik dan pembelajaran mendalam. Disini kita dapat melakukan penambangan data graf dalam hal pengambilan sampel graf, partisi graf, kompresi graf, pengelompokan graf, dan pencarian graf.

Kami dapat menskalakan pembelajaran mesin dengan metode pengambilan sampel data yang dapat mengatasi dimensi data dan granularitas data untuk multiprosesor dan pemrosesan batch paralel yang memalukan melalui tensor dan grafik. Pekerjaan yang terkait adalah mempelajari metode pengambilan sampel seperti undersampling, oversampling, ketidakpastian sampling, reservoir sampling, sampling struktural, dll. Solusi big data akan melibatkan operasi rekayasa data untuk caching, pengurutan, pengindeksan, hashing,

pengkodean, pencarian, partisi , pengambilan sampel dan pengambilan dalam model inkremental, model urutan, dan model ansambel untuk pembelajaran yang sensitif terhadap biaya dengan model grafis. Untuk analisis Sensitivitas pada data besar, kita dapat menganalisis metrik validasi prediksi yang menyesuaikan parameter jaringan neural dalam berdasarkan tren kesalahan klasifikasi dalam kumpulan data struktural.

Metrik validasi yang umum mencakup matriks konfusi, kurva perolehan presisi, kurva ROC, kurva peningkatan, dan statistik kappa. Di sini kami menemukan literatur tentang pembelajaran sekuens dan pembelajaran diskriminatif untuk memodelkan ekstraksi fitur dan residu regresi. Untuk pembelajaran representasi mendalam dari distribusi data tersebut, kita dapat menguraikan data historis menjadi pola terkini, sering, dan diawasi. Di sini kita dapat bereksperimen dengan metode diskritisasi seperti jendela geser, pembengkokan waktu dinamis, dan metode frekuensi waktu seperti wavelet, dan shapelet. Kami kemudian dapat memperlakukan distribusi data sebagai vektor 1D atau tensor 2D dalam pembelajaran mendalam untuk memperluas pemfilteran kolaboratif pada masukan pengguna akhir ke dalam kubus data yang bertindak sebagai struktur data untuk pembelajaran metrik jarak jauh.

Kami juga dapat menentukan struktur data sinopsis pada tensor dan grafik untuk mendapatkan fitur pembelajaran mesin. Struktur data sinopsis akan membantu pencarian kesamaan dan pembelajaran metrik dalam analisis jaringan yang kompleks. Dalam konteks ini kita dapat mengeksplorasi kausalitas dan stasioneritas rantai Markov dengan prinsip maksimalisasi ekspektasi dan panjang deskripsi minimum untuk inferensi statistik. Hasil analisis dapat diterapkan dalam tugas penambangan data seperti pengelompokan, klasifikasi, dan analisis asosiasi. Kita dapat memperluasnya ke dalam pembelajaran fitur untuk prediksi terstruktur, deteksi perubahan, penambangan peristiwa, dan penambangan pola dengan pembelajaran mendalam. Di sini fitur yang dipelajari dapat berupa salah satu fitur sampel, fitur yang dibangun, fitur yang diekstraksi, fitur yang disimpulkan, dan fitur prediktif.

Dalam hal estimasi parameter pemodelan, parameter regularisasi akan melakukan reduksi dimensi, sedangkan parameter pembelajaran melakukan klasifikasi prediktif dan regresi. Menggabungkan semua parameter ini dalam model data mining akan memungkinkan kita melakukan analisis sensitivitas model untuk sampel data yang berbeda. Minimalkan kesalahan pada berbagai jenis parameter dapat dimodelkan sebagai fungsi kerugian dalam model klasifikasi dan fungsi biaya dalam model optimasi untuk pembelajaran mendalam adversarial teoretis permainan. Jaringan saraf dalam yang relevan mencakup model berbasis fitur dan model berbasis memori. Pilihan antara jaringan saraf dalam untuk penambangan data ditentukan oleh metode pengujian hipotesis statistik dalam metode analisis data. Metode tersebut meliputi estimasi kemungkinan maksimum, uji hipotesis sekuensial, metode shift invarian, mesin vektor pendukung, dan metode dekomposisi tensor.

Grasedyck dkk. menghasilkan tinjauan literatur tentang teknik perkiraan tensor peringkat rendah dalam komputasi ilmiah. Penekanan khusus diberikan pada tensor yang disebabkan oleh diskritisasi fungsi multivariat yang mewakili solusi persamaan diferensial parsial berdimensi tinggi. Tensor tingkat tinggi menderita kutukan dimensi. Jadi harus didekati dengan skema kompresi seperti dekomposisi tensor dalam teknik tensor peringkat rendah.

Teknik tersebut memiliki beberapa penerapan seperti dalam solusi aproksimasi integral multidimensi, konvolusi multidimensi, persamaan diferensial parsial, persamaan Schrodinger, jaringan automata stokastik, keuangan komputasi, regresi multivariat, analisis wavelet, dan pembelajaran mendalam adversarial.

Beberapa dekomposisi tensor tingkat rendah yang populer untuk meningkatkan ketahanan pembelajaran adversarial dan efisiensi pembelajaran mendalam adalah dekomposisi CP, dekomposisi Tucker, dekomposisi rangkaian tensor, dan jaringan tensor. Algoritme komputasi yang menemukan dekomposisi meliputi metode iteratif yang dikombinasikan dengan pemotongan, algoritma berbasis optimasi, algoritma diskritisasi, algoritma dinamik, pendekatan peringkat-1 berturut-turut, dan pendekatan kotak hitam. Mereka dapat digunakan untuk memodifikasi, meningkatkan, dan mempelajari konvergensi prosedur pencarian dan pengoptimalan kuadrat terkecil bergantian dalam pembelajaran mendalam adversarial dengan musuh variasional.

Dengan demikian mereka dapat menghasilkan fungsi pembayaran manipulasi berbiaya rendah untuk skenario serangan yang jarang. Dalam konteks ini, kita dapat memperoleh model klasifikasi yang kuat dalam kerangka pembelajaran adversarial teoretis permainan. Kita dapat melakukan studi tentang fungsi biaya adversarial yang ada sehubungan dengan batasan ketahanan dan anggaran privasi dalam model pembelajaran representasi renggang untuk pembelajaran adversarial. Di sini kita juga dapat merumuskan ekstensi penambangan data dari pembelajaran manipulasi saya ke dalam penambangan web, analisis deret waktu, navigasi otonom sistem cyber-fisik dan manipulasinya, pengenalan pola multimedia, dan analisis keamanan jaringan.

Pemodelan teoretis permainan seperti itu berguna dalam pembelajaran mendalam yang bermusuhan atas kumpulan data multimodal, pengawasan yang lemah, berisik, jarang, tertaut, streaming, dan multi-terstruktur. Kita dapat menerapkan dinamika pengambilan sampel yang dihasilkan ke dalam penambangan data yang menjaga privasi dan pemrosesan sinyal fuzzy dari pola pencocokan yang berisik, jarang, dan lembut sebagai penyematan fitur dalam keamanan siber. Solusi keamanan siber dapat disediakan sebagai layanan cloud yang mengusulkan orkestrasi keamanan dalam arsitektur berorientasi layanan. Paradigma pembelajaran mesin yang relevan mencakup pembelajaran tambahan, pembelajaran online, pembelajaran penguatan, dan pembelajaran utilitas pada data aliran. Pekerjaan terkait adalah penambangan aliran data dengan informasi distribusi kelas dan biaya untuk fitur, anomali, hal baru, perubahan, dan komunitas di aliran.

Nouy memberikan survei metode tensor tingkat rendah untuk memperkirakan fungsi yang dinyatakan sebagai tensor dua tingkat untuk fungsi bernilai vektor, atau tensor tingkat tinggi untuk fungsi multivariat. Perkiraan peringkat rendah pada fungsi bernilai vektor dihitung dengan metode proyeksi berdasarkan sampel fungsi atau persamaan yang dipenuhi oleh fungsi tersebut. Dalam fungsi multivariat, perkiraan peringkat rendah sesuai dengan metode pengurangan pesanan model. Hal tersebut mencakup basis tereduksi, dekomposisi ortogonal yang tepat, subruang Krylov, metode pemotongan seimbang, dan dekomposisi

umum yang tepat. Model tersebut dapat diterapkan dalam analisis sensitivitas, kuantifikasi ketidakpastian, dan optimasi non-linier dalam pembelajaran adversarial teoretis permainan.

Metode pengurangan urutan model dapat bertindak sebagai metode perkiraan renggang yang memilih kamus fungsi yang memanfaatkan informasi sebelumnya tentang manipulasi pada manifold peringkat rendah. Bergantung pada kompleksitas komputasi metode ini dalam menangani kutukan dimensi, telah berkembang beberapa gagasan tentang peringkat dalam perkiraan peringkat rendah dari fungsi multivariat yang dapat diterapkan pada pembelajaran mesin. Perkiraan peringkat rendah dapat diperoleh dengan metode penyelesaian tensor yang merekonstruksi tensor dengan meminimalkan fungsi kerugian kuadrat terkecil. Pendekatan ganda mereka dapat menghasilkan regularisasi teoritis permainan dari masalah minimalisasi peringkat. Pertanyaan yang menantang mengenai kompleksitas komputasi dari pendekatan tersebut adalah jumlah sampel yang diperlukan untuk rekonstruksi stabil dari perkiraan peringkat rendah. Metode pemotongan peringkat rendah secara sistematis dapat membatasi penyimpanan dan kompleksitas komputasi dalam operasi aljabar.

Mereka memerlukan solusi masalah optimasi pada manifold berdimensi rendah dengan konstruksi serakah dari perkiraan peringkat rendah. Algoritme manipulasi yang dihasilkan dapat dianalisis sebagai versi tidak tepat yang menggabungkan gangguan data dalam pembelajaran mesin. Mengkarakterisasi pembelajaran adversarial teoretis permainan dalam kelas aproksimasi tersebut menghasilkan kelas fungsi kerugian adversarial dengan tingkat konvergensi aljabar atau eksponensial. Di sini pembelajaran kamus menghasilkan kamus elemen filter untuk merekonstruksi representasi distribusi data manipulasi yang sangat berlebihan dengan model pengkodean yang jarang dalam masalah optimasi & inferensi berbasis data. Sebagai paradigma pembelajaran mesin komputasi, pembelajaran kamus dapat menganalisis pembuatan fitur multimodal dan masalah pengoptimalan multivariat. Untuk mengkarakterisasi sinyal data dalam data eksperimen, kita juga dapat memperkirakan eksponen spektrum Lyapunov dan jaringan penarik dengan model pembelajaran mendalam dan proses analisis prediktif.

Kita dapat mengeksplorasi teknik pemrosesan sinyal spektral untuk menggambarkan dinamika kompleks dalam interaksi teoretis permainan sebagai fitur yang diekstraksi pada distribusi data mendasar yang menghasilkan dan memvalidasi manipulasi. Selanjutnya, kita dapat merekonstruksi persamaan diferensial yang mendasari model teoritis permainan sebagai sistem dinamis dari distribusi data pelatihan. Analisis lanskap pengoptimalan distribusi data dengan struktur pengindeksan tensor pada data statis dan dinamis dapat mengurangi biaya komunikasi dan meningkatkan penyeimbangan beban dalam sistem memori terdistribusi untuk pemrosesan paralel yang memalukan dari pembelajaran mendalam manipulasi teoretis permainan.

Pembaruan berulang terhadap perkiraan peringkat rendah dapat diperoleh dari trade-off antara kemampuan belajar dan ketahanan pembelajaran mendalam yang diawasi. Kita dapat mengkarakterisasi masalah diskriminasi keamanan siber dengan adanya gangguan manipulasi (adversarial noise) dalam bentuk serangkaian poin kuat di mana representasi data

menunjukkan jenis mekanisme mitigasi kesalahan spesifik masalah dalam desain pengklasifikasi. Pengaturan seperti itu juga akan memungkinkan untuk mengeksplorasi berbagai pilihan pengkodean variasional dari batas-batas keputusan yang dapat dipelajari dalam pemodelan teoretis permainan.

Filisbino dkk. memodelkan database gambar multidimensi dengan dekomposisi tensor. Komponen tensor yang diperingkat bertindak sebagai teknik reduksi dimensi. Mereka dapat digunakan untuk memperkirakan struktur kovarians suatu database dengan analisis subruang secara bersamaan. Mereka juga dapat menghitung bobot diskriminan dari pemisahan hyperplanes di dalamnya melalui analisis komponen utama diskriminan. Reduksi dimensi non-linier seperti itu menggeneralisasi pembelajaran mesin manipulasi dalam teknik reduksi dimensi linier seperti analisis komponen utama (PCA), analisis diskriminan linier (LDA), dan penyelesaian penskalaan multidimensi (MDS) untuk kriteria pengoptimalan linier. Mereka diklasifikasikan sebagai metode pembelajaran subruang dalam pembelajaran mesin.

Representasi tensor untuk gambar dapat diusulkan dengan metode pembelajaran subruang seperti metode dekomposisi nilai tunggal (SVD), analisis subruang konkuren (CSA), analisis komponen independen multilinear (MICA), analisis komponen utama multilinear (MPCA), analisis diskriminan tensor (TDA), dan dekomposisi tensor peringkat satu. Representasi gambar tersebut dapat diterapkan dalam pengenalan wajah dan gaya berjalan, pengenalan nomor digital, pemrosesan sinyal, analisis konten, dan deteksi anomali dalam penambangan data. Tugas analitik adalah mengidentifikasi “arah” yang paling diskriminan dalam analisis tensor yang diterapkan pada tugas klasifikasi tertentu. Ini adalah metode pemeringkatan fitur dalam subruang yang diproyeksikan untuk klasifikasi yang mengidentifikasi “arah” yang paling diskriminan dibandingkan fitur dengan variansi tertinggi dalam sampel data. Model pembelajaran mesin memeriksa keselarasan komponen utama tensor yang diberi peringkat dengan memisahkan arah hyperplane yang ditentukan oleh bobot diskriminan yang sesuai. Metode kernel seperti mesin vektor pendukung dapat membuat arah hyperplane.

Srebro dkk. memperkirakan matriks target dengan matriks peringkat rendah berbobot. Prosedur pemaksimalan harapan (EM) membuat parameter perkiraan yang mungkin tidak memiliki solusi bentuk tertutup. Matriks peringkat rendah tertimbang digunakan dalam pelatihan model faktor linier, regresi logistik, dan model kebisingan campuran Gaussians. Dekomposisi nilai tunggal (SVD) adalah salah satu pendekatan yang meminimalkan norma Frobenius ke matriks target. Pembobotan perkiraan dipengaruhi oleh gangguan yang merugikan sehingga menghasilkan rekonstruksi yang lebih baik dari struktur probabilistik yang mendasari dan distribusi statistik dalam data. Bobot juga dapat muncul dari batasan pada perkiraan yang dikodekan sebagai fitur dengan tingkat kepentingan yang berbeda-beda.

Hal ini juga dapat disebabkan oleh varians noise dan bias algoritmik dalam data pelatihan. Di sini pemodelan teori permainan dapat digunakan untuk mengoptimalkan perkiraan dan pembobotan ketika model kebisingan yang terkait dengan elemen matriks tidak diketahui. Hasilnya dapat dibandingkan dengan metode optimasi bergantian yang

sebanding. Metode optimasi bolak-balik memandang masalah perkiraan peringkat rendah tertimbang sebagai masalah kemungkinan maksimum dengan nilai yang hilang. Bobot matriks target dipetakan ke konfigurasi 0/1 di mana elemen yang diamati memiliki bobot 1 dan elemen yang hilang memiliki bobot 0. Biaya tertimbang suatu matriks setara dengan log-likelihood dari elemen yang diamati.

Algoritme EM memperbarui matriks parameter pada langkah ekspektasi untuk memaksimalkan kemungkinan log yang diharapkan dari matriks data di mana nilai yang hilang diperhitungkan sesuai dengan distribusi yang ditentukan oleh perkiraan kemungkinan log saat ini. Pada langkah maksimalisasi, matriks data diestimasi ulang sebagai perkiraan peringkat rendah tertimbang berdasarkan data. Sistem probabilistik seperti itu dapat diperluas ke beberapa matriks target dalam kerangka pembelajaran EM. Kemungkinan maksimum kemudian diperkirakan berdasarkan perkiraan peringkat rendah dari rata-ratanya jika matriks target dapat diamati sepenuhnya.

Jika beberapa matriks target tidak sepenuhnya diobservasi, algoritma EM dapat digunakan untuk mengisi nilai yang hilang dalam matriks target yang selanjutnya diestimasi sebagai pendekatan peringkat rendah. Pembaruan berulang pada matriks target dan bobot dapat didasarkan pada batasan variasi dalam memperkirakan kemungkinan log. Oleh karena itu, kita dapat menggabungkan iterasi perkiraan peringkat rendah berbobot dan iterasi terikat variasional sambil tetap memastikan konvergensi untuk keduanya.

Tsourakakis meningkatkan dekomposisi Tucker untuk menganalisis data multi-aspek dan mengekstrak faktor laten. Algoritme pengambilan sampel baru menghitung dekomposisi aliran tensor jika tensor tidak sesuai dengan memori yang tersedia. Dekomposisi Tucker dirumuskan sebagai masalah optimasi non-linier. Hal ini diselesaikan dengan algoritma optimasi Alternated Least Squares (ALS) yang mahal secara komputasi. Prosedur ALS dipercepat dengan algoritma acak yang memilih kolom berdasarkan distribusi probabilitas yang bias untuk dekomposisi tensor.

Mereka dapat diartikan sebagai generalisasi dari metode perkiraan tingkat rendah. Lebih jauh lagi, algoritme acak dapat menerima pemrosesan paralel yang memalukan pada aliran tensor. Perkiraan peringkat rendah tersebut mewakili bagian signifikan secara statistik dari data pelatihan yang diperoleh dari proses di dunia nyata. Mereka memiliki aplikasi dalam tugas penambangan data seperti deteksi anomali jaringan. Di sini outlier terdeteksi relatif terhadap subruang yang direntang oleh komponen utama dalam data pelatihan.

Zou dkk. mengusulkan analisis komponen utama renggang (SPCA) di mana jaring elastis menghasilkan komponen utama yang dimodifikasi dengan pembebanan renggang. Analisis komponen utama dilakukan sebagai masalah optimasi tipe regresi. SPCA tersebut memiliki aplikasi dalam klasifikasi kode pos tulisan tangan, pengenalan wajah manusia, analisis data ekspresi gen, dan analisis data multivariat. Richtarik dkk. membandingkan delapan formulasi pengoptimalan berbeda untuk SPCA dan implementasi paralelnya yang efisien pada multicore, GPU, dan cluster.

Formulasi yang kuat menggunakan fungsi tujuan yang merupakan fungsi dari matriks kovarians. Metode maksimalisasi bergantian adalah algoritma optimasi. Ini mengukur varians

data menggunakan norma L1 dan L2. Anandkumar dkk. mengusulkan dekomposisi tensor yang kuat menjadi komponen berperingkat rendah dan jarang. Metode yang diusulkan melakukan pendakian gradien pada bentuk variasi teregulasi dari masalah vektor eigen. Tujuan yang diatur memenuhi sifat konveksitas dan kehalusan untuk pengoptimalan. Momen empiris dalam probabilistik direpresentasikan sebagai tensor momen tingkat tinggi yang akan didekomposisi.

Kemudian korupsi momen diasumsikan terjadi akibat manipulasi atau bias sistematik dalam memperkirakan momen. Hasil eksperimen dibandingkan dengan matriks kuat PCA pada tensor pipih dan irisan matriks tensor. Mereka memiliki aplikasi dalam denoising gambar dan video, pembelajaran multitask, dan pembelajaran model variabel laten yang kuat. Romano dkk. menganalisis ketahanan pengklasifikasi terhadap gangguan manipulasi dengan menggunakan teori representasi renggang. Batasan diperoleh berdasarkan kinerja properti dan struktur pembelajar musuh dalam regresi dan klasifikasi.

Batasan tersebut ditunjukkan sebagai fungsi dari ketersebaran sinyal dan karakteristik filter/kamus/bobot pada sinyal yang masuk. Mereka mengungkap model data yang mengatur sensitivitas terhadap serangan musuh. Mekanisme regularisasi manipulasi berdasarkan solusi yang jarang dan kamus yang tidak koheren diusulkan untuk meningkatkan stabilitas pembelajar yang kuat dalam menghadapi kebisingan manipulasi. Hubungan sifat intrinsik sinyal dengan keberhasilan tugas klasifikasi dieksplorasi sebagai model generatif. Stabilitas model klasifikasi dipelajari dalam pengaturan kelas biner dan jamak.

Kreutz-Delgado dkk. mengembangkan algoritma berbasis data untuk pembelajaran kamus khusus domain. Mereka melakukan kemungkinan maksimum dan estimasi a posteriori maksimum. Prior diperoleh dari representasi sinyal lingkungan yang jarang yang disesuaikan dengan kamus sebagai konsep, fitur, dan kata. Dalam evaluasi eksperimental pembelajaran kamus yang diusulkan memiliki kinerja yang lebih baik dalam rasio signal-to-noise dibandingkan metode analisis komponen independen. Gambar yang dikodekan dengan kamus memiliki kompresi lebih tinggi (bit per piksel lebih sedikit) dan akurasi lebih tinggi (kesalahan kuadrat rata-rata lebih rendah).

Kamus ini memberikan representasi singkat dan ringkas untuk sebagian besar vektor sinyal yang mewakili secara statistik dalam lingkungan penghasil data. Struktur statistik dalam sinyal yang dihasilkan yang mencakup lingkungan pembelajaran diwakili dengan sekumpulan vektor basis yang mencakup berbagai sinyal bermakna berdimensi lebih rendah dalam kamus. Pembelajaran kamus memaksimalkan informasi timbal balik antara vektor basis dan sinyal yang dihasilkan. Memproyeksikan sinyal ke kamus menghasilkan pengurangan kebisingan dan kompresi data.

Masalah dekomposisi tensor dalam pembelajaran kamus adalah menghasilkan perkiraan tingkat rendah yang melengkapi kamus. Masalah representasi sinyal sebagai minimalisasi entropi menguraikan struktur statistik dalam distribusi data. Ini juga dapat dilihat sebagai generalisasi kuantisasi vektor. Model generatif stokastik dapat dikembangkan dalam pembelajaran mendalam untuk memecahkan masalah tersebut. Kombinasi teknik

pemaksimalan ekspektasi dan pendekatan variasional juga dapat digunakan dalam pembelajaran kamus.

Luedtke dkk. mengusulkan pembelajaran meta manipulasi Monte Carlo untuk membangun prosedur estimasi statistik yang optimal dalam masalah seperti estimasi titik dan estimasi interval. Permainan dua pemain dirumuskan antara Alam dan ahli statistik. Parameter jaringan saraf diperbarui berulang kali di seluruh interaksi game untuk mencapai representasi sampel terbatas yang diamati dalam eksperimen numerik. Oleh karena itu pembelajaran adversarial dapat dimasukkan ke dalam pendekatan frequentist dan Bayesian dalam mengukur kinerja pembelajaran mesin.

Dalam pendekatan frequentist, pembelajaran adversarial dapat menyelesaikan kinerja kasus terburuk dari penduga kemungkinan maksimum yang dinyatakan sebagai kriteria optimasi minimaxity. Dalam pendekatan Bayesian, pembelajaran adversarial dapat memperkirakan distribusi probabilitas posterior di mana optimasi minimax memperoleh prosedur Bayes dari campuran prior yang paling tidak disukai. Di sini risiko empiris maksimum dari suatu prosedur statistik dapat ditentukan dari distribusi yang paling tidak menguntungkan. Algoritme pembelajaran manipulasi Minimax memperbarui risiko tersebut secara berulang untuk meningkatkan model pembelajaran mesin. Prosedur statistik baru dapat dibangun untuk tugas penambangan data dengan cara yang hemat biaya menggunakan pembelajaran adversarial yang mendalam. Misalnya, Zhou dkk. menyajikan ansambel mesin vektor dengan relevansi yang jarang untuk pembelajaran manipulasi.

Selama pelatihan model, ia mampu memodelkan serangan manipulasi dengan parameter kernel. Konsep yang menyimpang ke arah parameter kernel meminimalkan kemungkinan titik data positif (berbahaya). Ini digunakan dalam pembelajaran bobot di mesin vektor relevansi. Di sini pemodelan teori permainan dapat dikatakan menyelesaikan masalah optimasi terbatas. Pengoptimalan tersebut dapat dibandingkan dengan metode non-teori permainan yang membuat asumsi tentang distribusi data yang rusak, sumber daya komputasi yang tersedia, dan pengetahuan musuh tentang model pembelajaran mesin yang ditargetkan.

Di sini ansambel mesin vektor relevansi (RVM) bertindak sebagai model berparameter linier jarang untuk pembelajaran manipulasi. RVM memiliki prioritas atas bobot yang akan diestimasi, dinyatakan sebagai sekumpulan hyperparameter yang terkait dengan bobot. Titik data pelatihan yang terkait dengan bobot bukan nol disebut vektor relevansi. Yin dkk. membatasi biaya pembuatan manipulasi dengan serangan fitur yang jarang dalam permainan jumlah bukan nol dengan musuh yang dianggarkan. Permainan non-zero sum memecahkan masalah regresi yang kuat.

Gemulla dkk. memfaktorkan matriks besar dalam algoritma optimasi stokastik berulang yang memperluas penurunan gradien. Perkiraan peringkat rendah dihasilkan dengan meminimalkan fungsi kerugian yang mengukur perbedaan antara matriks masukan asli dan produk dari faktor-faktor yang dikembalikan oleh algoritma faktorisasi. Kerugian strata didefinisikan pada kerugian yang dihitung pada setiap strata yang menyatakan matriks masukan sebagai gabungan potongan-potongan. Kriteria konvergensi dipelajari dengan mengacu pada teori pendekatan stokastik dan teori proses regeneratif. Varian penurunan

gradien dikhususkan untuk algoritme faktorisasi matriks yang dapat didistribusikan sepenuhnya dan dijalankan pada kumpulan data skala web. Oleh karena itu, faktorisasi matriks peringkat rendah sangat berguna dalam analisis data besar yang melibatkan kumpulan data besar di Internet.

Tugas analitik tersebut menemukan dan mengukur interaksi antara dua entitas tertentu dalam "data diadik" yang ditemukan dalam aplikasi seperti deteksi topik, pencarian kata kunci, dan personalisasi berita. Kerugian pelatihan dapat diatur dengan beberapa metode faktorisasi yang sesuai untuk platform pemrosesan terdistribusi yang sedang berkembang. Dia dkk. mengusulkan mesin faktorisasi saraf untuk prediktor kategoris dengan fitur biner yang sangat jarang. NFM memodelkan interaksi fitur yang mewakili struktur non-linier dan kompleks dalam data dunia nyata. Jaringan neural dalam mampu memodelkan interaksi fitur tingkat tinggi sebagai tensor tingkat rendah.

Mereka dapat menambah dan menggabungkan fitur kombinatorial yang menggabungkan beberapa variabel prediktor dalam rekayasa fitur pembelajaran mesin. Model pembelajaran mendalam mampu menggeneralisasi kombinasi fitur yang tidak terlihat dengan menyematkan fitur renggang berdimensi tinggi ke dalam ruang laten berdimensi rendah. NFM memiliki aplikasi sebagai metode penyematan untuk prediksi data renggang dalam periklanan online, pengambilan mikroblog, dan ekstraksi hubungan terbuka. Mereka dapat mempelajari interaksi fitur non-linier dengan menyematkan vektor fitur laten ke dalam berbagai arsitektur jaringan saraf dalam yang dibangun untuk meningkatkan kemampuan pembelajaran dan generalisasi.

Oleh karena itu, model ini lebih baik daripada model linier dalam mempelajari interaksi fitur seperti FM Tingkat Tinggi dan Mesin Eksponensial. Dalam tugas klasifikasi, regresi, dan pemeringkatan, mereka dapat mengatur fungsi kerugian seperti kerugian engsel, kerugian log, kerugian peringkat yang dipersonalisasi berpasangan, dan kerugian margin maksimal yang kontras. Oleh karena itu, mesin faktorisasi cocok untuk memodelkan kardinalitas tinggi dan distribusi data manipulasi yang jarang teramati. Mereka dapat mewakili kumpulan data yang ditemukan dalam analisis teks dan sistem pemberi rekomendasi dengan perkiraan matriks atau tensor peringkat rendah.

Mereka memungkinkan berbagai tugas analisis bisnis yang melibatkan data yang jarang seperti rekomendasi dan prediksi. Dia dkk. mengusulkan peringkat personalisasi adversarial (APR) untuk meningkatkan kekuatan model pemberi rekomendasi. APR menyempurnakan metode pemeringkatan fitur yang digunakan dalam rekomendasi top-k dengan pelatihan adversarial. Peringkat terpersonalisasi Bayesian (BPR) diambil saat pelajar berpartisipasi dalam permainan minimax dengan musuh yang menciptakan gangguan manipulasi pada parameter model untuk memaksimalkan fungsi objektif BPR. Gangguan manipulasi diperoleh dari penyematan vektor pengguna dan item yang berkontribusi terhadap pemfilteran kolaboratif. Pelatihan manipulasi mampu meningkatkan kesalahan generalisasi peringkat yang dipersonalisasi dalam model pemberi rekomendasi yang kuat.

Daftar rekomendasi teratas dievaluasi dengan ukuran kinerja seperti rasio hit (HR), keuntungan kumulatif diskon yang dinormalisasi (NDCG). Pengujian signifikansi statistik

dalam daftar peringkat yang dipersonalisasi untuk rekomendasi multimedia dilakukan dengan uji-t berpasangan satu sampel. Membuat model pemberi rekomendasi tahan terhadap contoh-contoh yang merugikan menghasilkan fungsi prediktif yang kuat dan stabil yang meningkatkan kinerja generalisasi dalam pengambilan informasi. APR dapat dikombinasikan dengan NFM untuk mendukung skenario rekomendasi seperti rekomendasi berbasis sesi yang cold-start, sadar konteks. Ini memiliki aplikasi dalam tugas pengambilan informasi seperti pengambilan teks, pencarian web, menjawab pertanyaan, dan penyelesaian grafik pengetahuan.

Metode Distribusi Relatif dalam Adversarial Deep Learning

Dalam teori pembelajaran komputasi, teori pembelajaran distribusi adalah kerangka untuk mempelajari distribusi dari sampel. Hal ini dapat dieksploitasi dalam pembelajaran mendalam manipulasi teoritis game untuk merancang algoritma perkiraan yang menargetkan model pembelajaran mesin. Kami merangkum ide-ide yang relevan dalam pembelajaran metrik jarak jauh dan pembelajaran metrik mendalam. Goldberger dkk. mengusulkan metrik jarak yang dipelajari yang digunakan dalam aturan pemilihan tetangga stokastik dalam metode tetangga terdekat untuk klasifikasi. Merupakan representasi data tingkat rendah yang mampu mengurangi biaya penyimpanan dan pencarian dalam pembentukan tetangga terdekat. Performa yang ditinggalkan adalah ukuran evaluasi yang dioptimalkan pada set data pelatihan.

Metrik jarak menghasilkan matriks jarak pada dataset pelatihan yaitu matriks semidefinite positif simetris yang digunakan dalam perhitungan jarak Mahalanobis. Ini memperkirakan transformasi ruang masukan di mana klasifikasi tetangga terdekat berkinerja baik. Aturan pemilihan tetangga stokastik memberikan penugasan lembut pada tetangga dalam fungsi tujuan pembelajaran yang diawasi. Memaksimalkan fungsi tujuan pembelajaran yang diawasi setara dengan meminimalkan norma L1 antara distribusi kelas sebenarnya dalam data dasar dan distribusi kelas stokastik yang diinduksi pada kumpulan data pelatihan. Memaksimalkan tujuan sesuai dengan klasifikasi bebas kesalahan dari seluruh kumpulan data pelatihan. Metrik jarak peringkat rendah yang diusulkan sebanding dengan teknik reduksi dimensi seperti analisis faktor, analisis komponen utama, analisis komponen independen, analisis diskriminan linier, dan analisis komponen yang relevan. Ini memecahkan masalah optimasi yang terbatas tanpa membuat asumsi parametrik tentang struktur distribusi kelas dan batasan keputusan.

Chopra dkk. membangun metrik kesamaan yang dapat dilatih untuk aplikasi pengenalan dan verifikasi. Metrik kesamaan mempelajari fungsi untuk memetakan pola masukan ke dalam ruang target sedemikian rupa sehingga norma L1 dalam ruang target mendekati jarak semantik dalam ruang masukan. Fungsi pemetaan dirancang sebagai jaringan saraf konvolusional yang tahan terhadap distorsi geometris. Fungsi kerugian diskriminatif meminimalkan metrik kesamaan untuk database wajah dengan variabilitas tinggi dalam pose, pencahayaan, ekspresi, posisi, dan oklusi buatan. Fungsi kerugian diturunkan dari model berbasis energi (EBM). Dibandingkan dengan model generatif, EBM tidak perlu memperkirakan distribusi probabilitas yang dinormalisasi pada ruang masukan.

Pendekatan terhadap tugas pengenalan seperti ini cocok untuk kumpulan data yang jumlah kategorinya besar dan jumlah sampel per kategori sedikit.

Xing dkk. mengusulkan masalah pembelajaran metrik jarak atas (dis) informasi sisi hubungan serupa dalam titik data. Pembelajaran metrik jarak dibingkai sebagai masalah optimasi cembung dengan solusi yang efisien. Metrik yang dipelajari dilatih melalui ruang fitur penuh dari masukan, bukan penyematan fitur yang berasal dari kumpulan data pelatihan. Jadi ia lebih mudah menggeneralisasi data yang sebelumnya tidak terlihat. Evaluasi eksperimental dilakukan terhadap varian K-means seperti constrained K-means, K-means + metric, dan constrained K-means + metric.

Kamu dkk. mengusulkan pembelajaran metrik jarak tertentu misalnya dalam metode tetangga terdekat. Ini menetapkan beberapa metrik ke lokasi berbeda dalam data pelatihan. Subruang METric Spesifik Instance (ISMETS) yang diusulkan mencakup ruang metrik secara generatif. Ini menginduksi subruang metrik untuk setiap contoh dengan menyimpulkan ekspektasi atas basis metrik dengan cara Bayesian. Inferensi statistik dilakukan menurut kerangka Bayes variasional. Bagian posterior menunjukkan keunggulan interpretabilitas, efektivitas, dan ketahanan. Dalam analisis data multimodal, pembelajaran metrik jarak jauh sebanding dengan pemrograman cembung terbatas, dan pendekatan teori informasi seperti pemodelan entropi maksimum. Ini dapat memprediksi metrik jarak untuk contoh pengujian yang tidak terlihat secara induktif dan transduktif. Ini dapat menggabungkan teknik paralelisasi dan trik perkiraan.

Shen dkk. mengusulkan teknik berbasis peningkatan untuk mempelajari metrik jarak Mahalanobis kuadrat. Solusi pemrograman semidefinite diberikan pada boosting. Ini menyatakan matriks semidefinit positif sebagai kombinasi linier dari matriks jejak-satu peringkat-satu. Mereka bertindak sebagai pembelajar yang lemah dalam proses pembelajaran berbasis peningkatan yang efisien dan terukur. Pemrograman semidefinite yang diusulkan dapat menggabungkan berbagai jenis batasan untuk agregasi peringkat dalam fungsi kerugian klasifikasi dan regresi.

Pembelajaran metrik jarak jauh terkait erat dengan metode subruang seperti analisis komponen utama, analisis diskriminan linier, proyeksi pelestarian lokalitas, dan analisis komponen relevan. Mereka dapat diartikan sebagai proyeksi data dari ruang masukan ke ruang keluaran berdimensi lebih rendah sambil mempertahankan struktur lingkungan dari kumpulan data pelatihan dalam pengertian teori informasi. Di sini pembelajaran metrik jarak jauh yang diawasi menggunakan informasi sampingan yang disajikan sebagai batasan pada masalah optimasi. Algoritme pendekatan serakah jarang memecahkan masalah optimasi dalam prosedur optimasi seperti AdaBoost untuk pemrograman semidefinite.

Sriperumbudur dkk. menganalisis metrik probabilitas integral (IPM) sebagai fungsi jarak yang dapat diukur antara dua distribusi probabilitas. IPM merupakan generalisasi dari metrik jarak populer seperti divergensi KL, divergensi Φ, divergensi Hellinger, divergensi Renyi, metrik Kantorovich, metrik Fortet-Mourier, perbedaan Stein, jarak Lipschitz, jarak variasi total, jarak Fisher, dan jarak kernel. Estimator empirisnya berguna dalam pembelajaran mesin untuk menghitung jarak antara distribusi data pelatihan dan manipulasi. Dalam

klasifikasi biner, IPM dapat diterapkan pada risiko empiris dan optimasi pemulusan antara distribusi kondisi kelas. Oleh karena itu, buruknya kesesuaian statistik terhadap kumpulan data pelatihan dapat diukur dengan ukuran divergensi probabilitas seperti IPM. Ini memiliki aplikasi untuk pemilihan model dalam desain pengklasifikasi dan estimasi kepadatan dalam skenario serangan manipulasi.

Hal ini berimplikasi pada kriteria konvergensi pembelajaran adversarial teori permainan pada khususnya dan pembelajaran adversarial generatif pada umumnya. Pilihan yang tepat dari ukuran divergensi probabilitas memberikan statistik uji signifikansi statistik pada hipotesis alternatif untuk pelatihan adversarial, fungsi kerugian yang efisien untuk ditargetkan dalam pembelajaran adversarial, dan perilaku konvergensi untuk pemodelan teoritis permainan. Dalam konteks ini, matematika seputar "jarak", "metrik", dan "divergensi" antara kumpulan data adversarial dan kumpulan data pelatihan menarik untuk memodelkan distribusi relatif yang mematuhi pertidaksamaan segitiga dalam pembelajaran adversarial teoretis permainan.

Liu dkk. pembelajaran metrik transfer survei untuk menganalisis data multimodal dalam aplikasi multimedia di mana domain target berada dalam tugas klasifikasi dan pencarian untuk analisis data. Berbeda dengan algoritma pembelajaran metrik transfer, algoritma pembelajaran metrik jarak jauh mengandalkan informasi label di domain target untuk pelatihan model. Pembelajaran metrik transfer dapat menangani informasi label terbatas dengan pembelajaran multiview. Representasi fitur multimodal untuk prediksi dengan pembelajaran transfer memiliki aplikasi dalam multimedia seperti analisis sentimen, penambangan opini, deteksi penipuan, deteksi penipuan internet, dan pencarian produk online.

Tujuan pembelajaran transfer adalah meningkatkan kinerja pembelajaran untuk tugas/domain yang diminati dengan menerapkan pengetahuan/keterampilan yang dipelajari dari tugas/domain terkait. Di sini pembelajaran metrik transfer memungkinkan transfer pengetahuan dengan estimasi jarak untuk metrik target linier dan non-linier untuk memandu klasifikasi multimodal dan aplikasi pencarian multimedia di domain target. Beberapa sampel berlabel digunakan dalam kombinasi dengan sejumlah besar kumpulan data tidak berlabel dalam fungsi kerugian klasifikasi multimodal.

Selanjutnya analisis fungsi kerugian berdasarkan peringkat dilakukan untuk aplikasi pencarian multimedia. Fitur SIFT seperti kata visual, tekstur wavelet, dan tag tekstual diturunkan sebagai fitur multimodal. Minimisasi divergensi untuk komputasi jarak pada beberapa domain dikategorikan sebagai minimalisasi divergensi berbasis representasi, minimalisasi divergensi berbasis jarak, dan minimalisasi divergensi berbasis kernel. Masalah minimalisasi divergensi diselesaikan dengan metode optimasi seperti maksimalisasi korelasi kanonik, minimalisasi divergensi matriks Burg, minimalisasi divergensi Bregman, minimalisasi divergensi determinan log, dan minimalisasi divergensi Von Neumann.

Belle dkk. mensurvei kegunaan pembelajaran metrik jarak jauh dalam pembelajaran mesin, pengenalan pola, dan penambangan data. Pembelajaran metrik adalah bidang penelitian yang secara otomatis mempelajari metrik jarak dari data. Pembelajaran metrik

jarak jauh Mahalanobis yang diawasi adalah dasar untuk perbandingan dengan metrik yang dipelajari. Varian algoritme pembelajaran metrik mencakup pembelajaran metrik non-linier, pembelajaran kesamaan, pembelajaran jarak jauh edit, pembelajaran metrik lokal, pembelajaran metrik multitugas, dan pembelajaran metrik semi-supervisi. Dalam konteks pembelajaran adversarial, pembelajaran metrik memungkinkan kita memperoleh jaminan generalisasi terhadap performa model pembelajaran mesin.

Kulis dkk. memberikan survei lain tentang penyetelan metrik jarak yang dipelajari ke tugas tertentu dalam analisis data dengan cara yang diawasi. Pembelajaran metrik yang diawasi didasarkan pada pelabelan informasi mengenai jarak data yang diubah. Hal ini menjadi perhatian khusus dalam menskalakan analisis data ke ruang fitur berdimensi tinggi dalam visi komputer, pengambilan gambar, pengenalan wajah, estimasi pose, analisis teks, analisis musik, analisis program, dan multimedia. Pembelajaran metrik memiliki ekstensi dalam regresi non-linier, pemeringkatan fitur, pengurangan dimensi, pengindeksan basis data, dan adaptasi domain. Jaringan pembelajaran mendalam memiliki peran penting dalam pengembangan metode pembelajaran metrik.

Hoffer dkk. mengusulkan model pembelajaran mendalam jaringan triplet untuk mempelajari representasi yang berguna melalui perbandingan jarak. Hal ini diterapkan dalam pembelajaran pemeringkatan dalam pengambilan informasi gambar. Fungsi kesamaan disebabkan oleh penyematan metrik norma untuk kumpulan data berlabel kelas jamak. Jaringan dalam adalah fungsi penyematannya. Ia menemukan jarak L2 antara masukan dari dua label dan representasi tertanam dari masukan label ketiga yang bertindak sebagai label referensi. Arsitektur jaringan saraf memungkinkan tugas analitik ini dinyatakan sebagai masalah klasifikasi dua kelas dengan tujuan fungsi kerugian adalah mempelajari penyematan metrik yang mengukur kedekatan dengan label referensi. Algoritme propagasi balik memperbarui model pembelajaran ini. Model ini mempelajari ukuran komparatif dibandingkan label kelas antara distribusi data yang diberi label. Mekanisme pembelajaran ini dapat dimanfaatkan untuk mengklasifikasikan sumber data baru yang labelnya tidak diketahui.

Chen dkk. mengusulkan jaringan manipulasi generatif berbasis metrik diskriminatif (DMGAN) yang menggunakan metode berbasis probabilitas untuk menghasilkan sampel mirip nyata dalam tugas sintesis gambar. Generator dilatih untuk menghasilkan sampel realistis dengan mengurangi jarak antara sampel nyata dan sampel yang dihasilkan. Diskriminator bertindak sebagai ekstraktor fitur yang mempelajari kerugian diskriminatif yang dibatasi oleh kerugian yang mempertahankan identitas. Kerugian diskriminatif memaksimalkan jarak antara sampel asli dan palsu dalam ruang fitur. Kerugian yang menjaga identitas menghitung jarak antara sampel dan pusatnya. Pusat-pusat tersebut diperbarui selama pelatihan GAN. Ini memetakan sampel yang dihasilkan ke dalam ruang fitur laten yang digunakan untuk memberi label pada sampel.

Dengan demikian DMGAN memulihkan distribusi implisit dari data sebenarnya. Ia mempelajari fitur-fitur representatif dalam ruang yang diubah. Kerugian yang mempertahankan identitas yang diusulkan dapat dibandingkan dengan kerugian triplet dan

kerugian kontrastif yang mempelajari variasi intrakelas dengan membatasi jarak antar sampelnya. Dengan demikian GAN dapat ditingkatkan dari perspektif pembelajaran metrik yang mendalam. GAN tersebut memiliki aplikasi dalam pembuatan gambar, resolusi super gambar, terjemahan gambar-ke-gambar, deteksi objek, dan pengenalan wajah. Dengan memperoleh sinyal propagasi balik melalui proses kompetitif, GAN tidak memerlukan komputasi probabilistik yang rumit seperti mesin Boltzmann dalam dan jaringan stokastik generatif.

Nowozin dkk. menafsirkan GAN sebagai pengambilan sampel saraf generatif di mana model probabilistik menerapkan pengambilan sampel. Model probabilistik tersebut menghasilkan sampel dari vektor masukan acak yang distribusi probabilitasnya ditentukan oleh bobot jaringan saraf. GAN dapat menghasilkan sampel tetapi tidak dapat menghitung kemungkinannya. Dari perspektif memperkirakan kemungkinan, f-GAN yang diusulkan menggeneralisasi metode pelatihan manipulasi di GAN menjadi proses estimasi divergensi variasional. f-GAN menggunakan f-divergence untuk melatih sampler saraf generatif. F-divergence dapat diganti dengan berbagai pilihan fungsi divergensi sehingga mengakibatkan perubahan kompleksitas pelatihan dan kualitas model generatif yang diperoleh. Ambang batas keputusan digunakan untuk mengklasifikasikan sampel generator.

Minimisasi divergensi variasional yang diusulkan dapat melakukan pengambilan sampel, estimasi, dan evaluasi kemungkinan dengan GAN. Hal ini sebanding dengan kombinasi jaringan kepadatan campuran dengan jaringan saraf berulang untuk menghasilkan model generatif teks tulisan tangan. Ini meningkatkan model probabilistik yang ada untuk pembelajaran mendalam seperti penaksir kepadatan autoregresif saraf bernilai nyata, model probabilistik difusi, dan estimasi kontrastif kebisingan. Ini dapat dikombinasikan dengan VAE untuk inferensi yang efisien. Ini dapat memperluas tujuan optimasi seperti perbedaan rata-rata maksimum kernel dengan metrik variasi total, jarak Wasserstein, dan jarak Kolmogorov.

Fedus dkk. memandang keseimbangan GAN sebagai keseimbangan Nash daripada minimalisasi divergensi antara distribusi pelatihan dan distribusi model. Jadi pemodelan teori permainan tentang keseimbangan GAN terbukti meningkatkan GAN minimax dalam hal kualitas dan keragaman sampel. Fungsi biaya adversarial digabungkan dengan tujuan minmax permainan sebagai fungsi regularisasi yang tidak jenuh sehingga sampel yang dihasilkan dihasilkan dengan kemungkinan besar menjadi nyata. Hukuman gradien pada manifold data yang dianalisis dari perspektif minimalisasi penyesalan dipilih sebagai tujuan regularisasi yang tidak jenuh. Algoritme tanpa penyesalan memperkirakan diskriminator di GAN bersifat linier di sekitar manifold data.

Lintasan menuju ekuilibrium Nash tidak sesuai dengan minimalisasi divergensi informasi secara bertahap. Sebaliknya dinamika pelatihan GAN mengoptimalkan metrik jarak berbeda yang diatur oleh fungsi biaya yang berlawanan. Oleh karena itu, kita dapat mengukur perbedaan informasi antara representasi minimal data pelatihan dan penyematan fitur data adversarial dengan fungsi biaya adversarial berbasis pembelajaran metrik yang mendalam. Kami juga dapat menerapkan distribusi sebelumnya pada faktor laten untuk pembuatan data yang koheren dalam pembelajaran generatif.

Bojanowski dkk. memperkenalkan Generative Latent Optimization (GLO) untuk melatih generator konvolusional mendalam menggunakan kerugian rekonstruksi. GLO adalah alternatif skema optimasi adversarial di GAN. GLO memungkinkan interpolasi linier dalam ruang kebisingan menjadi interpolasi semantik dalam ruang gambar, memungkinkan aritmatika linier dalam ruang kebisingan, dan memprediksi gambar target dari vektor kebisingan yang dapat dipelajari. Dalam evaluasi eksperimental, GLO dibandingkan dengan analisis komponen utama (PCA), autoencoder variasional (VAE), dan GAN. Keruntuhan mode di GAN diselidiki dengan kriteria rekonstruksi.

Bauso dkk. merumuskan permainan yang kuat secara distribusi menggunakan f-divergence dalam permainan multipemain antara skenario distribusi pelatihan dan distribusi manipulasi. Setiap pemain harus menghadapi distribusi kasus terburuk yang disebut distribusi adversarial. Algoritme pembelajaran Bregman mempercepat penghitungan kesetimbangan kuat. Skenario pembelajaran adversarial dipilih secara alami dan diasumsikan sebagai pemain virtual yang menyelesaikan fungsi tujuan non-cembung dan non-cekung. Sebuah teori trialitas diusulkan untuk pengurangan dimensi permainan yang kuat. Algoritme gerombolan memperkirakan penyelesaian gradien yang diharapkan untuk manipulasi.

Kamath dkk. mempelajari fungsi kerugian dalam masalah perkiraan distribusi dalam pembelajaran statistik di mana distribusi diperkirakan dari sampelnya. Dalam aplikasi kompresi, divergensi Kullback-Leibler direkomendasikan sebagai fungsi kerugian yang relevan. Dalam aplikasi klasifikasi, rugi-rugi L1 dan L2 direkomendasikan sebagai fungsi kerugian yang relevan. Dalam pembelajaran generatif, f-divergensi direkomendasikan sebagai fungsi kerugian yang relevan. Di sini kerugian kumulatif minmax untuk fungsi kerugian tertentu dan pencapaian estimator optimal memiliki kepentingan praktis dalam melatih model pembelajaran mesin.

Sugiyama dkk. membahas perkiraan dua distribusi probabilitas dari sampel mereka. Ini adalah masalah yang berimplikasi pada statistik, teori informasi, dan pembelajaran mesin. Divergensi Kullback-Leibler model estimasi kemungkinan maksimum dibandingkan oleh penulis dengan divergensi Pearson, jarak L2 untuk efisiensi, ketahanan, dan stabilitas. Di sini jarak yang tepat harus memenuhi pertidaksamaan segitiga yang merupakan perpanjangan teorema Pythagoras ke berbagai ruang metrik geometri. Mereka tidak boleh sensitif terhadap outlier. Jumlahnya tidak boleh tidak stabil. Mereka harus memiliki fungsi rasio kepadatan relatif yang terbatas dan efisien secara komputasi.

Penulis mensurvei beberapa aplikasi analisis data yang memanfaatkan langkah-langkah divergensi seperti deteksi titik perubahan, deteksi objek yang menonjol, dan estimasi keseimbangan kelas dalam beberapa tugas penambangan data seperti ekstraksi fitur, pengelompokan, analisis komponen independen, pembelajaran fitur kausal, analisis komponen independen, dan analisis ketergantungan kanonik. Perkiraan divergensi langsung yang dikombinasikan dengan reduksi dimensi dikatakan sebagai strategi yang lebih baik dalam eksperimen daripada estimasi kepadatan distribusi sampel yang naif. Perbedaan antara jarak statistik dan divergensi informasi adalah pengaruhnya terhadap kriteria konvergensi dalam rangkaian distribusi probabilitas yang dipelajari yang diperkirakan dengan

model generatif dan metode variasional. Divergensi yang dioptimalkan biasanya terputus-putus sehubungan dengan parameter generator. Jadi cara-cara baru untuk memperkirakan secara praktis fungsi rasio kepadatan relatif minimum dan tertinggi harus dirancang dalam pembelajaran mendalam manipulasi berdasarkan geometri metrik, probabilitas terapan, dan statistik.

5.5 MEKANISME PERTAHANAN DALAM PEMBELAJARAN MESIN PERMAINAN STRATEGIS

Zhang dkk. mengusulkan mekanisme pertahanan dalam manipulasi data yang merugikan pada waktu pengujian. Serangan penghindaran tersebut mengaburkan konten email spam dan mengeksploitasi kode yang tertanam dalam sampel malware dan paket jaringan. Keamanan pengklasifikasi ditemukan memburuk dengan pemilihan fitur. Jadi properti keamanan pemilihan fitur diselidiki terhadap serangan penghindaran. Implementasi berbasis wrapper diusulkan untuk menggabungkan strategi manipulasi dalam deteksi spam dan malware dengan pemilihan fitur sadar-musuh dalam pengklasifikasi dengan fungsi diskriminan linier dan non-linier.

Dalam tugas penambangan data yang sensitif terhadap keamanan, memilih subset fitur yang relevan akan meningkatkan kinerja generalisasi pengklasifikasi, mengurangi kompleksitas pembelajaran komputasi, dan memungkinkan pemahaman yang lebih baik tentang detail pemodelan. Selama proses pemilihan fitur, keamanan pengklasifikasi dimodelkan sebagai istilah regularisasi untuk dioptimalkan bersama dengan kemampuan generalisasi pengklasifikasi. Jarak antara sampel yang dimanipulasi dan sampel yang sah serta batasan pada pengklasifikasi dan representasi fitur digunakan untuk mengembangkan algoritma manipulasi yang efisien dalam pengaturan serangan blackbox.

Evaluasi keamanan pengklasifikasi manipulasi dilakukan terhadap serangan dengan kekuatan yang meningkat. Hal ini berkorelasi dengan figur kebajikan yang disebut kekerasan penghindaran. Di sini norma L1 mendorong ketersebaran dalam manipulasi, berbeda dengan norma L2. Penanggulangan serangan penghindaran secara eksplisit memasukkan pengetahuan tentang manipulasi ke dalam algoritma pembelajaran. Hal ini mencakup algoritme pembelajaran manipulasi teoritis permainan, model probabilistik dari strategi serangan yang dihipotesiskan, kombinasi pengklasifikasi yang lebih lemah dalam beberapa sistem pengklasifikasi, dan sanitasi data berdasarkan statistik yang kuat.

Biggio dkk. melakukan evaluasi keamanan mesin vektor dukungan (SVM) yang tergabung dalam sistem keamanan dunia nyata. Mereka terlibat dalam perlombaan senjata dalam domain aplikasi keamanan seperti deteksi malware, deteksi intrusi, dan pemfilteran spam dengan kompleksitas dan keterpaparan yang semakin meningkat. Oleh karena itu, pola pembelajaran mesin harus dimasukkan ke dalam aplikasi keamanan untuk melengkapi deteksi berbasis tanda tangan tradisional pada sampel tanpa filter dan serangan tidak populer.

Pola serangan tersebut dikategorikan sebagai serangan keracunan yang menyesatkan algoritma pembelajaran, serangan penghindaran yang menghindari deteksi pada waktu penerapan, dan pelanggaran privasi yang memperoleh informasi tentang detail pemodelan.

Di sini musuh memanipulasi data untuk mengeksploitasi kerentanan dalam algoritma pembelajaran yang membuat asumsi stasioneritas dalam teknik berbasis evaluasi kinerja seperti validasi silang, bootstrapping, dan minimalisasi risiko empiris. Desain SVM yang sadar akan musuh dirancang sebagai teknik penanggulangan. Kerangka privasi diferensial diusulkan sebagai tindakan penanggulangan serangan privasi.

Survei perlombaan senjata antara musuh dan pengklasifikasi dirinci dalam kaitannya dengan fitur pembelajaran mesin yang dieksploitasi dalam pengklasifikasi spam gambar dan pendeteksi outlier di jaringan komputer. Dalam konteks ini, masalah pembelajaran adversarial dapat dianggap sebagai perlombaan senjata proaktif di mana pengklasifikasi mengantisipasi pergerakan musuh. Evaluasi keamanan terhadap solusi pembelajaran adversarial dilakukan dengan kriteria yang bergantung pada aplikasi yang direpresentasikan dalam skenario serangan yang dihipotesiskan. Dampak serangan dievaluasi dalam bentuk fungsi kerugian terbatas pada kumpulan data berlabel yang bertindak sebagai statistik gabungan dari data sensitif.

Biggio dkk. memandang serangan keracunan sebagai jenis outlier dalam data pelatihan. Ansambel pengantongan berbobot kemudian diusulkan sebagai tindakan penanggulangan terhadap serangan keracunan. Dengan demikian, masalah perancangan pengklasifikasi yang kuat dirumuskan dalam rangka memitigasi sampel outlier dalam data pelatihan dengan mengurangi komponen varians estimasi atau kesalahan klasifikasi atau regresi. Dengan demikian statistik yang kuat dapat mengurangi efek serangan keracunan pada data pelatihan. Aplikasi keamanan siber ditampilkan dalam pemfilteran spam dan deteksi intrusi.

Deteksi intrusi difokuskan pada aplikasi web di lingkungan yang kritis terhadap keamanan seperti sistem medis, keuangan, militer, dan administrasi. Tujuan musuh mengirimkan pertanyaan berbahaya adalah untuk mengakses informasi rahasia atau menyebabkan penolakan layanan. Biggio dkk. merancang pengklasifikasi yang kuat dengan menghasilkan distribusi data untuk tugas klasifikasi manipulasi dari model estimasi kemungkinan maksimum. Aplikasi keamanan siber ada pada verifikasi identitas biometrik dan pemfilteran spam. Label kelas untuk pembelajaran yang diawasi adalah berbahaya (M) atau sah (L) untuk mengklasifikasikan pengguna yang mengakses sistem komputer sebagai “asli” (L) atau “penipu” (M). Dasar deteksi spam untuk pengklasifikasi teks Naive Bayes adalah penyisipan kata yang baik (GWI) dan kebingungan kata yang buruk (BWO) terhadap filter spam berbasis teks.

Penanggulangan memodifikasi algoritma klasifikasi dalam fase pelatihannya. Garis dasar dalam ciri-ciri biometrik adalah serangan spoof terhadap sistem biometrik multimodal untuk verifikasi identitas. Mereka dianggap sebagai serangan integritas eksplorasi. Kinerja dievaluasi menggunakan kurva karakteristik operasi penerima (ROC), yang menunjukkan persentase pengguna asli yang diterima (tingkat penerimaan asli, GAR) sebagai fungsi dari persentase penipu yang diterima (tingkat penerimaan palsu, FAR), untuk semua nilai ambang batas keputusan.

Dekel dkk. menyajikan pengklasifikasi kuat yang memperkirakan masalah pembelajaran dengan pemrograman linier yang dianalisis untuk memberikan batas risiko statistik pada perbedaan antara distribusi data pelatihan dan klasifikasi. Pembelajaran statistik dalam perceptron kemudian membahas varian masalah pembelajaran online. Skema regularisasi A L digunakan untuk menyeimbangkan ketersebaran dan kepadatan dalam pembelajaran pengklasifikasi yang rentan terhadap gangguan yang merusak fitur. Meminimalkan risiko empiris dirumuskan sebagai masalah optimasi kombinatorial.

Pengklasifikasi online dibatasi pada hyper-cube untuk mengontrol kompleksitasnya. Contoh pengujian dirusak oleh musuh yang serakah. Pengorbanan komputasi terlihat pada ketahanan pelatihan pengklasifikasi pada data yang jarang dan padat. Xu dkk. menunjukkan kesetaraan antara mesin vektor dukungan yang diatur (SVM) dan formulasi optimasi yang kuat. Kekokohan dikatakan sebagai alasan untuk kinerja generalisasi dalam SVM untuk kelas kumpulan ketidakpastian tipe non-kotak. Teori optimasi yang kuat digunakan untuk memotivasi konstruksi istilah regularisasi dalam pembelajaran mesin untuk pengaturan pembelajaran non-i.i.d.

Sampel pengujian dianggap sebagai gangguan terhadap sampel pelatihan. Formulasi SVM seperti itu didasarkan pada pengklasifikasi yang dibatasi oleh peluang. Dalam meminimalkan batas atas kesalahan klasifikasi yang diharapkan, secara matematis setara dengan formulasi minmax dari masalah optimasi dalam pemodelan teoretis permainan. Tampilan pengoptimalan SVM yang kuat dapat memperoleh batasan kompleksitas sampel untuk kelas algoritma klasifikasi yang luas. Namun, ketahanan dalam ruang fitur yang dijamin oleh proses regularisasi tidak menjamin ketahanan dalam ruang observasi karena pemetaan fitur yang “tidak mulus” di kernel tertentu.

Demontis dkk. menghubungkan perkembangan optimasi yang kuat dengan ketersebaran, regularisasi, dan keamanan pengklasifikasi linier. Pengklasifikasi linier digunakan dalam sistem tertanam dan perangkat seluler karena interpretasi keputusannya, waktu pemrosesan yang rendah, dan kebutuhan memori yang kecil. Ketersebaran bobot fitur ternyata mempunyai efek yang diinginkan tidak hanya pada biaya pemrosesan tetapi juga keamanan pengklasifikasi linier. Pengoptimalan yang kuat terbukti berdampak pada regularisasi pengklasifikasi di mana serangan penghindaran dianggap sebagai bentuk gangguan manipulasi. Oleh karena itu, masalah pembelajaran mesin adversarial adalah memilih pengatur yang optimal terhadap berbagai jenis noise adversarial.

Biaya manipulasi dalam memodifikasi data dinyatakan dalam norma L1 yang menghasilkan serangan jarang yang sebanding dengan jarak e antara sampel asli dan sampel yang dimodifikasi dalam ruang Euclidean. Strategi serangannya adalah meminimalkan fungsi diskriminan pengklasifikasi sehingga sampel berbahaya diklasifikasikan sebagai sampel sah dengan tingkat keyakinan tinggi. Masalah optimasi yang kuat didefinisikan dalam bentuk gangguan terbatas pada data pelatihan dan kumpulan ketidakpastian terkait termasuk bola L1. Masalah pembelajarannya adalah meminimalkan kerugian diskriminatif untuk masalah klasifikasi dua kelas dalam kasus terburuk, gangguan terbatas pada data pelatihan. Pengatur yang diusulkan ditunjukkan untuk mengimbangi ketersebaran fitur dengan biaya komputasi

evaluasi keamanan. Aplikasi keamanan siber ditampilkan untuk klasifikasi digit tulisan tangan, pemfilteran spam, dan deteksi malware PDF.

Feng dkk. mengusulkan algoritma regresi logistik yang kuat. Hal ini kuat terhadap outlier manipulasi dalam matriks kovariat yang rusak. Prosedur pemrograman linier sederhana mempelajari parameter regresi logistik dalam masalah klasifikasi biner. Hal ini dibandingkan dengan metode pembobotan ulang berulang untuk mengoptimalkan regresi logistik. Adversarial outlier bersifat arbitrer, tidak terbatas, dan tidak berasal dari distribusi tertentu. Mereka menyimpang dari estimasi parameter dalam regresi logistik untuk menurunkan kinerjanya.

Kurva regresi yang dihasilkan jauh dari kenyataan sebenarnya. Prediksi yang diberi label pada inlier salah. Fungsi kerugian dalam regresi logistik adalah kerugian 0-1, kerugian engsel, kerugian eksponensial, dan kerugian logistik. Daripada melakukan inferensi kemungkinan, penduga kuat dan parameter regresinya dapat diusulkan dengan memperkirakan secara kuat statistik korelasi linier seperti matriks kovariat. Batasan teoritis dapat diperoleh dari batas risiko empiris dan populasi pada regresi logistik. Skala penduga kuat yang diusulkan untuk masalah besar berisi sampel pelatihan yang rusak dengan cara komputasi yang efisien.

Barreno dkk. memberikan taksonomi serangan serta pertahanan pada algoritma dan sistem pembelajaran mesin. Model pembelajaran mesin menawarkan manfaat dengan dilatih tentang perbedaan baru antara distribusi data normal (yang diketahui baik) dan serangan (yang diketahui buruk). Ruang hipotesis atau kelas fungsi untuk model pembelajaran mesin yang diawasi tersebut terdiri dari tabel pencarian, fungsi linier, polinomial, fungsi Boolean, dan jaringan saraf. Pengorbanan teori pembelajaran adalah antara penjelasan dan generalisasi. Mereka memiliki aplikasi dalam pemfilteran email spam, deteksi kesalahan, deteksi intrusi, deteksi virus, layanan web, sistem agen, dan pemantauan cluster yang harus bersaing dengan pola data yang berubah secara dinamis. Algoritme pembelajaran yang lebih rumit berlatih dengan aliran titik-titik tak berlabel yang berkelanjutan dalam paradigma pembelajaran online untuk pembelajaran semi-supervisi.

Ruang fitur untuk membangun batas keputusan antara titik data normal dan serangan adalah ruang metrik pada jarak antar titik. Regularisasi diusulkan sebagai mekanisme pertahanan dalam serangan kausatif, sedangkan pengacakan direkomendasikan dalam serangan eksplorasi. Memperlancar solusi pembelajar menghilangkan kompleksitas yang dapat dieksploitasi oleh musuh. Serangan yang ditargetkan lebih sensitif terhadap variasi batasan keputusan. Pemrosesan awal pada distribusi sebelumnya dapat menyandikan pengetahuan domain pada garis dasar untuk estimasi pembelajar. Pengacakan dalam posisi batas keputusan disarankan sebagai mekanisme pertahanan dalam serangan yang ditargetkan. Algoritme verifikasi tanda air digital yang tersedia untuk umum juga harus menangani sensitivitas serangan yang ditargetkan. Sejumlah besar titik data yang salah klasifikasi merupakan indikasi serangan kausatif. Serangan penyebab dapat dimitigasi dengan kumpulan data pengujian yang terdiri dari intrusi yang diketahui. Serangan eksplorasi dicirikan oleh kelompok besar yang tiba-tiba di dekat batas keputusan.

Mereka dapat dideteksi dengan menjalankan algoritma pengelompokan pada dataset pelatihan pengklasifikasi. Mendeteksi serangan semacam itu memberikan informasi kepada pelajar tentang kemampuan musuh yang dapat digunakan dalam mekanisme pertahanan. Model teoritis permainan seperti permainan penipuan memformalkan informasi seperti manipulasi data manipulasi yang dibangun oleh setiap pemain. Mereka melibatkan sebagian informasi untuk setiap pemain dan pengaruhnya terhadap informasi yang dilihat oleh pemain lain. Informasi tersebut dikodekan sebagai distribusi probabilitas dan status diskritisasi dalam fungsi pembayaran manipulasi di mana biaya dikaitkan untuk mengubah fitur di titik serangan bagi musuh dan mengukur setiap fitur dalam data untuk pelajar. Dalam permainan yang lebih rumit, pembelajar dapat mengacaukan perkiraan lawan mengenai keadaan pembelajar. Tujuan pembelajar adalah untuk mengizinkan intrusi "honeypot" untuk mengelabui musuh dan mencegahnya mempelajari batasan keputusan.

Dalam kasus seperti ini, peran pembelajar dan lawan menjadi terbalik. Biaya tindakan penanggulangannya dikatakan bergantung pada dampak data yang sah terhadap proses pembelajaran. Seorang pelajar yang memasukkan informasi sebelumnya kehilangan kemampuan beradaptasi terhadap data baru. Pada saat yang sama pembelajar yang menampung informasi dari data pelatihan menjadi lebih rentan terhadap serangan. Oleh karena itu kita harus mempertimbangkan faktor keamanan dan kerahasiaan batasan keputusan dalam proses pembelajaran untuk pelatihan ulang. Hal ini mencakup trade-off antara jumlah data pelatihan untuk pelajar dalam pelatihan dan kerahasiaan pelajar yang dihasilkan dalam penerapan.

Hubungan antara pelatihan ulang bertingkat dan pengetahuan domain musuh merupakan masalah terbuka untuk penelitian. Kita dapat menghasilkan batasan teori informasi atas informasi yang diperoleh musuh dengan mengamati perilaku pelajar pada titik data tertentu. Bergantung pada rincian klasifikasi pelajar dalam situasi realistis, batas kepercayaan dapat diserang dengan kekuatan prediksinya. Mekanisme pembobotan adaptif dalam teori permainan seperti algoritma agregasi dan algoritma mayoritas tertimbang dapat menggabungkan saran dari sekumpulan pakar untuk memprediksi rangkaian interaksi teoretis permainan yang memiliki kinerja pembelajaran yang sebanding dengan pakar terbaik dalam rangkaian yang dipilih secara musuh. Sistem kendali adalah alternatif teori permainan yang dapat diterapkan pada sistem pakar berorientasi pencarian dalam komando dan kendali militer.

Biggio dkk. melakukan survei sistem kompleks untuk pengenalan pola dalam situasi manipulasi. Di sini musuh dapat merancang serangan untuk mengeksploitasi langkah-langkah pra-pemrosesan seperti kesalahan penguraian atau kerentanan rumit dalam algoritma pembelajaran seperti penolakan layanan, deteksi yang hilang, dan sampel atau peristiwa berbahaya. Kemudian paradigma keamanan reaktif dan proaktif dapat dimanfaatkan oleh pelajar untuk meningkatkan keamanannya sesuai desain. Analisis sistem dan komponen desain untuk pengenalan pola dipusatkan pada pengumpulan data data pelatihan dan kebenaran dasarnya, pemrosesan awal untuk mengekstrak komponen struktural, ekstraksi fitur pada sampel yang diurai, pemilihan fitur dengan atau tanpa pengawasan manusia,

algoritma pembelajaran untuk membangun pengklasifikasi dari suatu kumpulan data berlabel, dan aturan keputusan untuk menetapkan label pada sampel pengujian masukan berdasarkan strategi ambang batas pada skor pengklasifikasi.

Komponen-komponen tersebut dapat ditempatkan di lokasi fisik yang berbeda sehingga musuh dapat menargetkan saluran komunikasi yang memerlukan otentikasi jarak jauh dan protokol keamanan untuk pengawasan manusia. Permukaan serangan dapat dibangun untuk semua kerentanan operasional ini. Hal ini berguna untuk membuat hipotesis statistik mengenai tujuan dan pengetahuan musuh tentang sistem target dan kemampuannya untuk memanipulasi data yang menyebabkan kegagalan sistem. Umpan balik terhadap keputusan pengklasifikasi dapat meningkatkan pengetahuan musuh tentang bagaimana sistem pengenalan pola diterapkan, di mana sistem tersebut diterapkan, dan kapan sistem tersebut beroperasi. Jadi, mata rantai keamanan terlemah dalam sistem pengenalan pola tidak selalu pada komponen pembelajaran atau klasifikasi. Pembelajaran dengan invarian seperti itu penting untuk pendekatan minimax dengan kompleksitas komputasi tinggi dalam pembelajaran mendalam adversarial teoretis permainan. Ini memiliki aplikasi dalam otentikasi pengguna, visi komputer dan forensik, analisis sentimen, dan segmentasi pasar.

Barreno dkk. menyajikan taksonomi serangan terhadap sistem pembelajaran mesin. Hal ini dapat digunakan untuk menyusun biaya bagi musuh dan pembelajar dalam membangun sistem pembelajaran yang aman dan tahan terhadap serangan. Analisis kesalahan pelajar mengasumsikan pengaturan klasifikasi biner. Perluasannya ke pengaturan multi-label tidaklah mudah. Dalam prosedur pelatihan ulang, pengklasifikasi menyisipkan pelatihan dengan evaluasi. Pelatihan ulang semacam itu dapat dianalisis dalam kerangka minimalisasi risiko empiris (yang diatur). Risiko empiris dihitung sebagai perkiraan kerugian dari fungsi kerugian yang mendekati biaya sebenarnya. Regularisasi risiko empiris mencegah overfitting pada data pelatihan dengan gagasan kompleksitas hipotesis pada distribusi data non-stasioner.

Model ancaman yang ditimbulkan oleh musuh dinyatakan dalam tujuan/insentif penyerang dan kemampuan penyerang. Pilihan yang dibuat oleh musuh dan pembelajar disajikan sebagai strategi khusus domain dengan fungsi biaya terkait yang menilainya. Misalnya, model pembelajar mungkin merupakan mesin vektor pendukung dengan kernel yang dipilih, kerugian, regularisasi, dan rencana validasi silang dalam hipotesis pembelajaran. Musuh kemudian memilih prosedur statistik untuk menghasilkan distribusi data untuk mengevaluasi dan memvalidasi hipotesis pembelajaran. Prosedur statistik juga dapat memperlakukan pelajar sebagai oracle yang memberikan label untuk menanyakan contoh dalam serangan probing. Dengan probing, musuh mungkin menemukan titik data berbiaya tinggi bagi pembelajar. Di sini permainan one-shot meminimalkan biaya manipulasi ketika setiap gerakan terjadi, sementara permainan berulang meminimalkan total akumulasi biaya yang ditemukan dengan memainkan permainan yang diulang beberapa kali.

Penulis merangkum biaya tersebut dan pertimbangan praktisnya dalam beberapa model teoritis permainan untuk serangan ketersediaan kausatif dan serangan integritas eksplorasi. Mekanisme pertahanan mereka dalam pengklasifikasi manipulasi mengubah

fungsi kemungkinan pelajar sehingga dapat mengukur setiap fitur dengan biaya berbeda yang diketahui. Musuh kemudian bermain secara optimal melawan pengklasifikasi asli yang sensitif terhadap biaya. Di sini bidang penelitian statistik yang kuat dapat membandingkan prosedur kandidat untuk merancang prosedur untuk mencapai pembelajar yang kuat. Hal ini dapat digunakan untuk mengembangkan teori informasi untuk sistem pembelajaran aman yang dapat mengukur kebocoran informasi dalam jumlah bit. Hal ini juga dapat mengukur risiko empiris yang terkait dengan serangan saluran sampingan terhadap informasi yang bocor.

Mekanisme Pertahanan dalam Pembelajaran Mendalam Adversarial

Untuk menciptakan pembelajaran mesin yang tangguh dalam pendeteksian malware, Tong dkk. mengidentifikasi fitur-fitur yang dilestarikan yang tidak dapat dimodifikasi tanpa mengorbankan fungsi berbahaya. Mereka digunakan untuk membangun pertahanan yang berhasil melawan serangan penghindaran yang dapat direalisasikan. Ketahanan pembelajaran mesin kemudian digeneralisasikan ke beberapa serangan yang dapat direalisasikan untuk melakukan pengerasan model dengan ruang fitur yang memperhitungkan serangkaian serangan yang dapat direalisasikan dalam pengoptimalan yang kuat.

Kumpulan ekstraktor fitur dirancang untuk menghitung nilai vektor numerik dan label objek terkait untuk fitur dari entitas masukan terkait. Bergantung pada asumsi tentang algoritme pembelajaran dan model manipulasi, pertahanan penghindaran diklasifikasikan menjadi penalaran teori permainan, pengoptimalan yang kuat, dan pelatihan ulang manipulasi yang berulang. Generalisasi pertahanan penghindaran dievaluasi berdasarkan model ruang fitur serangan penghindaran yang dapat direalisasikan. Pengklasifikasi PDF berbasis struktur pada fitur biner dari properti struktural dalam file PDF serta pengklasifikasi PDF berbasis konten pada metadata dan konten PDF digunakan untuk membedakan antara kejadian jinak dan berbahaya.

Serangan penghindaran yang dapat direalisasikan dibuat dengan EvadeML yang memiliki akses kotak hitam ke pengklasifikasi, serangan mimikri yang memanipulasi file PDF berbahaya menggunakan injeksi konten agar menyerupai file PDF jinak, MalGAN untuk menghasilkan contoh malware, serangan mimikri terbalik untuk memasukkan muatan berbahaya ke dalam file jinak target, dan serangan khusus untuk mengganti entri dalam file PDF serangan dengan representasi heksadesimal yang mengaburkan tag untuk eksekusi kode dalam PDF. Pelatihan ulang manipulasi yang berulang dipilih sebagai mekanisme pertahanan untuk menghasilkan pengklasifikasi yang kuat.

Chaowei dkk. memanfaatkan informasi konteks spasial dalam segmentasi semantik untuk mendeteksi contoh manipulasi bahkan ketika berhadapan dengan musuh adaptif yang kuat. Hipotesis mekanisme pertahanan adalah bahwa contoh manipulasi dalam tugas pembelajaran mesin yang berbeda mengandung sifat statistik unik yang memberikan pemahaman mendalam tentang potensi mekanisme pertahanan. Dalam tugas segmentasi semantik, ini berarti memberikan label prediksi pada setiap piksel dalam gambar berdasarkan informasi kontekstual dalam lingkungan spasialnya. Target kebenaran dasar yang berlawanan kemudian ditentukan dalam kumpulan data video mengemudi otonom di dunia nyata. Contoh

manipulasi ditemukan tidak dapat ditransfer antara strategi deteksi dengan pembelajaran mendalam di antara berbagai tugas/skenario segmentasi.

Dia dkk. menyimpulkan bahwa pertahanan gabungan yang menggabungkan beberapa pertahanan lemah tidak menciptakan pertahanan yang kuat terhadap berbagai contoh manipulasi yang dibuat oleh musuh adaptif. Feiman dkk. mengusulkan untuk mendeteksi contoh manipulasi dari perkiraan kepadatan pada ruang fitur lapisan tersembunyi terakhir. Estimasi ketidakpastian Bayesian juga digunakan untuk mendeteksi sampel adversarial di wilayah dengan tingkat kepercayaan rendah pada ruang masukan. Sebaliknya, Raghunathan dkk. menghasilkan sertifikat ketahanan untuk jaringan saraf yang mekanisme pertahanannya didasarkan pada regularisasi dan pelatihan manipulasi.

Sertifikat memastikan kesalahan akibat serangan manipulasi dibatasi oleh nilai tertentu untuk berbagai contoh manipulasi. Di sini, pelatihan manipulasi dikatakan meminimalkan batas bawah kerugian terburuk karena pelatihan tersebut tidak dapat digeneralisasikan ke serangan baru yang dirancang untuk menyesatkan pengoptimal. Sertifikat kerugian adversarial merupakan relaksasi semidefinite pada pengoptimal yang dapat dihitung secara efisien. Hal ini kontras dengan batas atas pada kerugian manipulasi akibat spektral dan norma Frobenius. Pekerjaan terkait ditemukan dalam literatur teori kontrol tentang verifikasi ketahanan sistem dinamis. Fungsi Lyapunov dapat digunakan untuk memodelkan evolusi nilai aktivasi dari waktu ke waktu dalam jaringan saraf sebagai sistem dinamis yang berubah terhadap waktu sehingga batas kerugian adversarial dapat dipahami dalam kaitannya dengan bukti stabilitas pada lintasan sistem ini.

Metode sertifikasi mengenai stabilitas dan kinerja kelompok model dalam sistem pembelajaran mesin juga dapat memperoleh manfaat dari verifikasi keselamatan dan prosedur sintesis pengontrol seputar representasi data kuat yang digunakan dalam robotika untuk infrastruktur penting. Miyato dkk. mengusulkan metode regularisasi baru untuk pelatihan manipulasi tanpa overfitting. Kerugian manipulasi virtual ditentukan berdasarkan kekokohan distribusi label bersyarat di sekitar setiap titik data masukan. Istilah regularisasi diartikan sebagai pendistribusian terlebih dahulu tentang pengetahuan atau keyakinan yang apriori tentang model pembelajaran. Keyakinan khusus yang diambil dari hukum fisika adalah bahwa kondisi kelas posterior dari sistem pembelajaran mesin mulus sehubungan dengan masukan spasial dan/atau temporal.

Kelancaran distribusi lokal dari distribusi keluaran sehubungan dengan distribusi masukan didefinisikan sebagai ketahanan model distribusi berbasis divergensi terhadap arah manipulasi virtual. Arah manipulasi virtual diartikan sebagai arah paling anisotropik yang memberikan label “virtual” pada titik data yang tidak berlabel. Regularisasi kebisingan manipulasi yang dihasilkan meningkatkan kinerja generalisasi dalam tugas klasifikasi gambar semi-supervisi.

Papernot dkk. memperkenalkan distilasi defensif untuk mempertahankan jaringan saraf dalam dari sampel musuh. Ini adalah prosedur pelatihan untuk jaringan dalam menggunakan pengetahuan yang ditransfer dari jaringan dalam yang berbeda. Motivasi penyulingan adalah untuk mengurangi kompleksitas komputasi arsitektur pembelajaran

mendalam dengan mentransfer pengetahuan dari arsitektur yang lebih besar ke arsitektur yang lebih kecil sehingga pembelajaran mendalam dapat diterapkan pada perangkat cyber-fisik yang memiliki sumber daya terbatas. Distilasi defensif menerapkan ide ini untuk mengekstraksi pengetahuan dari jaringan saraf dalam guna meningkatkan ketahanannya.

Transfer pengetahuan digunakan untuk mengurangi amplitudo gradien jaringan dalam yang dieksploitasi oleh musuh. Oleh karena itu, model yang dilatih dengan distilasi defensif kurang sensitif terhadap sampel musuh. Sebagai tindakan pencegahan keamanan, hal ini menghasilkan model pengklasifikasi yang lebih halus dengan sifat generalisasi yang lebih baik. Distilasi defensif mengarah ke dua bagian selama pelatihan yang disebut estimasi sensitivitas arah dan pemilihan gangguan di mana tujuan manipulasi diasumsikan sebagai kesalahan klasifikasi sampel dari kelas sumber tertentu ke dalam kelas target yang berbeda.

Pengetahuan yang ditransfer tidak hanya terdiri dari parameter bobot yang dipelajari oleh deep net tetapi juga vektor probabilitas kelas terkode yang dihasilkan oleh jaringan selama pelatihan. Probabilitas kelas lunak lebih baik daripada label kelas keras karena label tersebut menyimpan informasi relatif tentang entropi kelas selain kelas yang benar dari setiap sampel. Informasi tentang probabilitas bersyarat kelas tersebut dapat digunakan untuk memandu konvergensi jaringan dalam menuju solusi pemodelan optimal yang meningkatkan ketahanan klasifikasi.

Untuk menghadapi serangan optimasi dengan tujuan dan pengoptimal baru, Papernot et al. memperluas lipatan dalam Papernot dkk. untuk menambahkan kelas outlier untuk mengurangi contoh manipulasi dan memberikan perkiraan ketidakpastian dalam jaringan saraf melalui inferensi stokastik. Dengan mentransfer pengetahuan dan ketidakpastian, penyulingan defensif yang diperluas tidak memerlukan pembela untuk menghasilkan contoh-contoh manipulasi menurut heuristik. Tramer dkk. memperkenalkan pelatihan manipulasi ansambel untuk menambah data pelatihan dengan gangguan manipulasi yang ditransfer dari model terlatih lainnya.

Memasukkan serangan blackbox dalam gangguan manipulasi tersebut secara signifikan meningkatkan kemampuan pengalihan contoh-contoh manipulasi. Mekanisme pertahanan seperti ini berguna dalam serangan multilangkah yang lebih mahal. Metode tanda gradien cepat (FGSM) dan variannya seperti metode kelas satu langkah yang paling tidak mungkin terjadi (Step-LL) dan serangan berulang (I-FGSM atau Iter-LL) digunakan untuk membuat contoh manipulasi. Baik musuh whitebox maupun blackbox digunakan untuk mengevaluasi perolehan kekuatan dalam strategi pertahanan. Oleh karena itu, pelatihan manipulasi ditingkatkan dengan memisahkan pembuatan contoh manipulasi dari pelatihan model. Pada saat yang sama, musuh interaktif juga diusulkan untuk memasukkan kueri pada fungsi prediksi model target dalam serangan mereka.

Wu dkk. mengusulkan kerangka kerja tetangga dekat yang sangat percaya diri untuk menggabungkan informasi kepercayaan prediksi dan pencarian tetangga terdekat untuk memperkuat ketahanan manipulasi. Meng dkk. mengusulkan kerangka MagNet untuk mekanisme pertahanan. Ini mencakup jaringan detektor terpisah dan jaringan reformer untuk mendeteksi contoh-contoh manipulasi. Jaringan detektor adalah pembuat enkode otomatis

untuk mempelajari kumpulan data dari contoh normal tanpa mengasumsikan proses stokastik tertentu untuk menghasilkannya. Mereka dilatih berdasarkan kriteria kerugian rekonstruksi yang memperkirakan jarak antara masukan dan berbagai contoh normal.

Jaringan reformer adalah autoencoder lain yang memindahkan contoh-contoh yang berlawanan ke kumpulan data dari contoh-contoh normal untuk mengklasifikasikannya dengan benar. Berdasarkan ide kriptografi, pertahanan melalui keragaman dianjurkan untuk memilih secara acak satu dari beberapa pertahanan pada saat run time dalam serangan kotak abu-abu. Carlini dkk. kemudian dapat membuat contoh manipulasi yang dapat ditransfer untuk MagNet dan eliminasi gangguan manipulasi GAN (APE-GAN). Berdasarkan metrik jarak, Carlini et al. juga berhasil membangun contoh-contoh manipulasi untuk jaringan yang disaring secara defensif.

Metzen dkk. menambah jaringan saraf dalam dengan jaringan detektor. Ia melakukan klasifikasi biner antara data asli dan data manipulasi untuk mendeteksi jenis musuh tertentu. Ini dapat bertindak sebagai metode untuk memperkuat detektor terhadap musuh yang dinamis. Cisse dkk. memperkenalkan jaringan Parseval untuk analisis empiris dan teoretis tentang kekuatan prediksi yang dibuat oleh jaringan dalam yang tunduk pada gangguan manipulasi. Mereka bertindak sebagai metode regularisasi dengan batasan ortonormalitas untuk mengurangi efek manipulasi.

Demikian pula, Gu dkk. mengusulkan Jaringan Kontraktif Dalam (DCN) yang bertindak sebagai hukuman kelancaran pada pelatihan manipulasi. DCN adalah perpanjangan dari autoencoder kontraktif (CAE) yang memiliki kemampuan untuk menghilangkan gangguan manipulasi. Jadi ide-ide dari denoising autoencoder (DAE), contractive autoencoder (CAE), dan marginalized denoising autoencoder (mDAE) memberikan kerangka kerja yang kuat untuk melatih jaringan saraf dalam dengan kriteria ketahanan yang disesuaikan dengan persepsi manusia. Sebaliknya, Kos dkk. membuat contoh manipulasi di ruang laten untuk model generatif mendalam seperti autoencoder variasional (VAE) dan jaringan manipulasi generatif (GAN).

Xiao dkk. mengusulkan AdvGAN untuk menghasilkan contoh manipulasi dengan jaringan manipulasi bersyarat dalam skenario serangan semi-kotak putih dan kotak hitam. Jin dkk. mengusulkan APE-GAN untuk bertahan melawan contoh-contoh manipulasi dalam skenario serangan kotak putih. Generator mengubah gangguan manipulasi dengan perubahan kecil pada contoh masukan. Diskriminator dioptimalkan untuk memisahkan contoh bersih dan contoh yang direkonstruksi tanpa gangguan yang merugikan. Fungsi kerugian diciptakan untuk membuat contoh manipulasi konsisten dengan keragaman data gambar asli. APE-GAN dapat dikombinasikan dengan mekanisme pertahanan lain seperti pelatihan ulang manipulasi.

Dengan asumsi hipotesis bahwa contoh manipulasi terletak pada wilayah distribusi pelatihan dengan probabilitas rendah, Song et al. merancang PixelDefend untuk memindahkan gambar berbahaya yang terganggu kembali ke distribusi yang terlihat di data pelatihan. Model generatif menghitung probabilitas semua gambar pelatihan. Kepadatan probabilitas seperti itu digunakan untuk menentukan peringkat contoh-contoh manipulasi

yang diciptakan oleh berbagai metode penyerangan. Masalah optimasi terbatas yang sulit diselesaikan dirumuskan untuk memurnikan contoh-contoh yang berlawanan. Hal ini diperkirakan dengan prosedur decoding serakah. Hasilnya dibandingkan dengan mekanisme pertahanan lain dalam literatur seperti pelatihan manipulasi, penghalusan label, dan pemerasan fitur.

Bojanowski dkk. memperkenalkan kerangka kerja Generative Latent Optimization (GLO) untuk melatih generator menggunakan kerugian rekonstruksi. Hal ini berguna dalam interpolasi antara sampel pelatihan dan contoh manipulasi. Hal ini juga memungkinkan aritmatika linier antara vektor kebisingan di ruang laten untuk mempelajari interpolasi contoh manipulasi tanpa memerlukan permainan manipulasi antara generator dan diskriminator. Generator kemudian menerjemahkan interpolasi linier pada ruang noise menjadi interpolasi semantik pada ruang gambar. Ruang kebisingan yang dapat dipelajari mampu menguraikan faktor-faktor non-linier variasi ruang gambar menjadi statistik linier.

Kyatham dkk. menggabungkan gangguan manipulasi ke ruang laten generatif yang diatur dan dikuantisasi untuk kemudian memetakannya ke kumpulan data sebenarnya. Mekanisme pertahanan berdasarkan autoencoder generatif kemudian mampu menghindari kelemahan mekanisme pertahanan terkait seperti perkiraan turunan dalam pelatihan manipulasi dan parameterisasi ulang ekspektasi dalam penyaringan manipulasi. Encoder laten mempertahankan jarak dalam ruang metrik pada manifold data dan laten. Hal ini memungkinkan eksplorasi stokastik dari lingkungan laten dari distribusi yang diketahui. Beberapa dekoder digunakan untuk dengan mudah menjelajahi sampel data di ruang laten dan memetakannya kembali ke titik data di data yang sah. Eksplorasi ruang laten dapat dilakukan dengan inferensi variasional yang dilakukan dengan representasi laten hierarki yang dipelajari dari data seperti dalam autoencoder variasi tangga dan Mask Adversarial Auto-Encoder.

Fawzi dkk. memberikan kerangka teoritis untuk menganalisis ketahanan pengklasifikasi yang terkena gangguan manipulasi. Batas atas ketahanan ditetapkan untuk pengklasifikasi linier dan kuadrat. Kekokohan dinyatakan sebagai ukuran keterbedaan antar kelas. Untuk pengklasifikasi linier, ini adalah jarak antar sarana kelas. Untuk pengklasifikasi kuadrat, ini adalah jarak antara matriks momen orde kedua dari kelas-kelas tersebut. Batasan kekokohan untuk pengklasifikasi ditetapkan secara independen dari algoritma pembelajaran dan mekanisme pertahanan. De Silva dkk. mengusulkan tindakan penanggulangan pembelajaran adversarial (CAL) yang sadar biaya terhadap serangan untuk merancang pengklasifikasi yang tahan terhadap serangan. Informasi struktur biaya serangan diperoleh dari analisis kerentanan sistem pembelajaran mesin.

Kerangka kerja CAL memproyeksikan contoh pengujian yang berpotensi dipalsukan ke dalam ruang vektor fitur yang sah dengan fungsi biaya serangan yang bertindak sebagai metrik jarak. Operator proyeksi dapat diartikan sebagai uji rasio kemungkinan yang digeneralisasi. Kerangka kerja CAL dapat diterapkan pada teknik klasifikasi apa pun. Insua dkk. memberikan analisis risiko manipulasi untuk klasifikasi manipulasi yang dapat dianggap sebagai alternatif kerangka teori permainan. Skenario serangan dibatasi pada serangan

eksplorasi dan serangan pelanggaran integritas. Pendekatan Bayesian terhadap klasifikasi adversarial digunakan untuk mengusulkan pengklasifikasi generatif. Pendekatan frequentist terhadap klasifikasi adversarial digunakan untuk mengusulkan pengklasifikasi yang diskriminatif.

Kegunaan yang diharapkan terkait dengan konsekuensi terburuk dari musuh adalah kriteria optimasi. Model sampel generatif dalam distribusi data adversarial menyebabkan kesulitan komputasi pada data berdimensi tinggi. Jadi contoh-contoh manipulasi diambil sampelnya secara algoritmik dari wilayah dengan kerugian manipulasi yang tinggi dengan istilah noise untuk memperhitungkan ketidakpastian pembela HAM terhadap model penyerang. Ketidakpastian penyerang terhadap model pembela disebabkan oleh pendekatan Bayesian yang dapat diskalakan terhadap pembelajaran mendalam. Secara umum, asumsi spesifik aplikasi direduksi seminimal mungkin sambil mempelajari utilitas dan probabilitas yang merugikan dalam paradigma Bayesian. Analisis risiko manipulasi seperti itu memungkinkan kita untuk menggabungkan metode Bayesian dengan pembelajaran mendalam manipulasi teoritis permainan.

Schmidt dkk. mempelajari kompleksitas sampel teori informasi dari pembelajaran kuat yang tidak bergantung pada algoritma pelatihan atau keluarga model. Karena sifat statistik dari pembelajaran mendalam, contoh-contoh manipulasi dikatakan terbukti terjadi dalam setiap pendekatan pembelajaran. Batas bawah ditetapkan pada kekerasan dan ketahanan dalam pengklasifikasi mendalam yang sesuai dengan musuh terbatas yang menerapkan pergeseran distribusi kasus terburuk. Musuh tidak adaptif terhadap pengaturan pengklasifikasi. Dengan demikian, kesenjangan yang jelas dapat ditunjukkan antara generalisasi yang kuat dan standar untuk kelas hipotesis dan kelas distribusi dalam pembelajaran mesin manipulasi.

Akibatnya, mekanisme pertahanan harus dibuat khusus untuk jenis musuh tertentu dan kumpulan data pelatihan tertentu. Menambahkan informasi sebelumnya tentang trade-off ketahanan ke dalam arsitektur model dapat membantu menciptakan pengklasifikasi yang kuat. Wong dkk. mengatasi kesenjangan ketahanan antara gangguan di dunia nyata dan kumpulan data yang umum terjadi pada pertahanan musuh dengan kumpulan gangguan yang dapat dipelajari. Generator bersyarat mendefinisikan kumpulan gangguan pada wilayah terbatas ruang laten. Generator bersyarat adalah autoencoder variasi bersyarat. Hal ini dapat menimbulkan gangguan pada kompleksitas dan skala yang berbeda mulai dari transformasi gambar dasar. Sebagai model ancaman, kumpulan gangguan pembelajaran dengan metrik kuantitatif secara empiris dan tersertifikasi kuat terhadap manipulasi, variasi, dan korupsi yang merugikan.

Seshia dkk. berpendapat perlunya spesifikasi semantik dan konteks sistem pembelajaran mesin di lingkungan dengan sumber daya terbatas. Spesifikasi semantik tingkat sistem tentang manipulasi dapat digunakan untuk menghasilkan tidak hanya label yang salah klasifikasi tetapi juga informasi semantik mengenai implikasi tingkat sistem. Pendekatan manipulasi semantik seperti ini berguna untuk aplikasi pembelajaran manipulasi dalam sistem tertanam yang melibatkan perangkat Internet of Things (IoT) dan sistem kontrol cyber-

fisik (CPCS). Kemudian komponen ketahanan manipulasi di sekitar spesifikasi tingkat sistem dalam algoritma pelatihan dapat dinyatakan sebagai ruang modifikasi semantik, fungsi kerugian semantik untuk pelatihan, dan augmentasi kumpulan data semantik.

Bahasa pemrograman probabilistik dapat digunakan untuk memandu pembelajaran manipulasi semantik dengan mewakili asumsi distribusi mengenai pembuatan data, inferensi, dan verifikasi. Metode algoritmik dapat digunakan dalam perancangan dan analisis sistem pembelajaran berbasis AML/A2I. Spesifikasi formal model pembelajaran mendalam dan sistem pembelajaran mesin juga dapat memperoleh manfaat dengan mengeksplorasi trade-off antara ketahanan semantik dan implementasi hemat sumber daya dalam desain sistem yang tahan terhadap kesalahan.

Rouhani dkk. mengusulkan kerangka kerja otomatis DeepFense untuk pelaksanaan pembelajaran mendalam yang efisien dan aman dalam aplikasi penting dan sensitif terhadap waktu seperti kendaraan tak berawak, drone, dan sistem pengawasan video. DeepFense memanfaatkan redundansi modular dalam desain bersama perangkat keras/perangkat lunak/algoritma untuk mencapai kinerja tepat waktu dalam pengaturan sumber daya terbatas. Deteksi sampel manipulasi online dievaluasi pada FPGA dan GPU. Setiap redundansi modular mempelajari fungsi kepadatan probabilitas dari titik data umum dan wilayah langka/berisiko terkait dengan pembelajaran kamus. Selain validasi performa pembelajaran mesin, performa sistem disesuaikan dengan latensi, konsumsi energi, dan jejak memori di sekitar penyediaan sumber daya perangkat keras yang mendasarinya.

Pertukaran kemampuan pembelajaran ditemukan tidak hanya antara kinerja sistem dan ketahanan terhadap persaingan, namun juga keterbatasan sumber daya dan keandalan model sistem pembelajaran mesin. Kekokohan pembelajaran mendalam yang bermusuhan harus sesuai dengan batasan pengoptimalan yang ditentukan pengguna dan/atau khusus perangkat keras. FPGA digunakan untuk memberikan paralelisme yang terperinci dan respons yang tepat waktu dalam mekanisme pertahanan musuh. Tolok ukur pembelajaran mendalam kemudian dipamerkan untuk melawan serangan manipulasi yang canggih. Rantai modul pembela Markov digunakan untuk memitigasi serangan musuh adaptif. DeepFense disajikan sebagai metode pembelajaran tanpa pengawasan untuk menghaluskan batasan keputusan guna menghilangkan variabel kebisingan yang merugikan. Hal ini terbukti meningkatkan ketahanan model pembelajaran mendalam dengan mempelajari kepadatan data manipulasi di ruang laten akibat serangan adaptif terhadap mekanisme pertahanan online yang ada.

Ghafouri dkk. mengusulkan regresi terawasi sebagai mekanisme pertahanan untuk mendeteksi manipulasi pembacaan sensor dalam sistem cyber-fisik (CPS). Interaksi antara pembela dan penyerang CPS dimodelkan sebagai permainan Stackelberg di mana pembela memilih ambang batas keputusan yang kira-kira optimal menggunakan regresi yang diawasi dan deteksi anomali manipulasi untuk meningkatkan ketahanan sistem pembelajaran mesin. Interaksi tersebut dapat digunakan untuk mengembangkan algoritma heuristik untuk detektor tangguh dengan pemodelan berbasis regresi dalam jaringan sensor. Detektor tangguh yang dihasilkan dapat digunakan untuk menentukan batasan kuat pada

kemungkinan manipulasi sensor dengan memanfaatkan hubungan antar pengukuran dari beberapa sensor.

Taran dkk. mengusulkan mekanisme pertahanan baru berdasarkan prinsip kriptografi Kerckhoffs kedua. Hal ini mengarah pada skenario serangan kotak abu-abu di mana musuh memiliki akses ke prediksi klasifikasi, data pelatihan/pengujian, dan label kelas untuk pembelajaran mesin tetapi tidak memiliki kunci rahasia kriptografi terkait yang mengenkripsi pengetahuan tentang parameter mekanisme pertahanan. Kunci rahasia menerapkan blok encoder pra-pemrosesan yang dapat diimplementasikan sebagai transformasi data-independen dalam berbagai cara karena musuh tidak dapat mendekripsi parameter mekanisme pertahanan pengklasifikasi dalam jumlah waktu yang wajar dalam sarana komputasi modern yang tersedia.

Mekanisme pertahanan kriptografi yang diusulkan dapat diintegrasikan dengan pertahanan melalui pelatihan ulang, pertahanan melalui deteksi dan penolakan, pertahanan melalui pemrosesan awal masukan, dan pertahanan melalui regenerasi. Jadi mengintegrasikan kriptografi asimetris dengan pembelajaran adversarial menghasilkan keuntungan informasi dari pembela/pelajar dibandingkan penyerang/musuh. Entropi kunci rahasia lebih tinggi daripada entropi sinyal musuh. Transformasi datanya tidak dapat dibedakan. Untuk lebih melindungi mekanisme pertahanan, arsitektur sistem pengklasifikasi diasumsikan tidak dapat diakses oleh musuh. Hal ini mengarah pada desain protokol pembelajaran dengan prinsip kriptografi dalam sistem pengenalan pola yang diterapkan pada server yang dilindungi atau perangkat atau chip khusus untuk aplikasi pembelajaran yang bermusuhan seperti watermarking digital, kriptanalisis, forensik digital, steganalisis, dan identifikasi perangkat.

Xu dkk. menyelidiki batasan teori informasi dari pembelajaran manipulasi. Ia memiliki aplikasi dalam domain solusi seperti visi komputer, pengawasan video, pemrosesan bahasa alami, pengenalan suara, dan keamanan siber. Hal ini juga berlaku untuk masalah pembelajaran seperti klasifikasi, regresi, penyematan fitur (dalam kata dan node), dan model generatif. Musuh seharusnya memiliki anggaran serangan terhadap kebisingan musuh yang dapat disuntikkan pada waktu pembelajaran. Anggaran serangan dinyatakan sebagai jarak statistik yang disebut jarak variasi total (TVD) antara distribusi data asli dan distribusi data yang diinjeksi noise. Pekerjaan terkait pada batasan statistik ketahanan adversarial mencakup batasan generalisasi untuk pembelajaran adversarial, sertifikasi ketahanan untuk inferensi, kemampuan pembelajaran PAC kelas VC yang kuat, dan analisis efek memasukkan kebisingan ke dalam jaringan pada waktu inferensi.

Jha dkk. membuat pemecah teori modulo kepuasan (SMT) dengan kombinasi pembelajaran yang dipandu Oracle dari contoh dan sintesis berbasis batasan dari komponen di perpustakaan pembelajaran mesin. Sintesis program otomatis untuk deobfuscation program berguna dalam verifikasi formal pembelajaran adversarial. Oracle validasi memeriksa apakah program pembelajaran mesin benar atau tidak berdasarkan persyaratan keamanan pembelajaran adversarial. Ini memiliki koneksi ke prosedur optimasi dalam teori pembelajaran komputasi dan program manipulasi bit. Dalam lingkungan manipulasi non-

stasioneritas, Lowe et al. mengeksplorasi metode pembelajaran penguatan mendalam untuk domain multi-agen.

Metode aktor-kritik dapat diadaptasi untuk pembelajaran adversarial untuk mempelajari kebijakan seputar interaksi teoretis permainan dalam pembelajaran mendalam adversarial melalui koordinasi multi-agen yang kompleks. Mereka memecahkan kebijakan multi-agen yang kuat dalam skenario serangan kooperatif dan kompetitif dari perilaku yang muncul dan kompleksitas agen yang berevolusi bersama. Dengan demikian pembelajaran penguatan dapat diterapkan pada lingkungan pembelajaran manipulasi dengan banyak musuh.

Li dkk. membuat pendeteksi malware yang tangguh untuk contoh manipulasi pada malware Android. Kombinasi autoencoder variasional (VAE) dan multi-layer perceptron (MLP) digunakan untuk merancang fungsi kerugian baru yang menguraikan fitur-fitur dari kelas malware yang berbeda. Ruang fitur malware Android direpresentasikan secara terpisah. Mekanisme pertahanan yang diusulkan menghitung metrik kesamaan antara contoh-contoh yang tidak berbahaya dan berbahaya sambil mempertahankan fungsionalitas yang berbahaya. Model klasifikasi akhir secara bersamaan melakukan pendeteksian malware dan pertahanan contoh manipulasi.

Hassan dkk. membahas perlindungan batas kepercayaan untuk memungkinkan hak istimewa akses pengguna di lingkungan Industrial Internet of Things (IIoT). Musuh dapat menggunakan teknik model skewing untuk menghasilkan contoh manipulasi pada permukaan serangan di jaringan IIoT. Generator data kooperatif berbasis downsampler-encoder digunakan untuk membuat contoh manipulasi di perangkat IIoT. Perangkat IIoT tersebut mencakup perangkat IoT seperti sensor, pengontrol logika yang dapat diprogram, aktuator, perangkat elektronik cerdas, dan sistem cyber-fisik (CPS) dalam operasi industri.

CPS mencakup subsistem dan proses untuk desain, infrastruktur, pemantauan dan pengendalian, penjadwalan, dan pemeliharaan rantai nilai analisis data untuk kontrol proses fisik yang tepat, manajemen kolaborasi sistem industri yang otonom, pengumpulan data produksi yang lebih murah, dan pemrosesan cerdas secara real-time . Kerentanan dan ancaman terhadap protokol, jaringan, sistem, dan layanan industri tersebut terbuka untuk dieksploitasi oleh musuh. Hal ini semakin diperburuk dengan adanya celah keamanan dalam sistem TI konvensional. Di sini mekanisme pertahanan dalam pembelajaran mendalam adversarial digunakan untuk menegakkan tujuan keamanan data IIoT seperti kerahasiaan, integritas, dan ketersediaan.

Penerapan lebih lanjut dari pembelajaran mendalam manipulasi diberikan oleh Abusnaina dkk. dan Martins dkk. Abusnaina dkk. menganalisis deteksi malware IoT dengan fitur berbasis grafik aliran kontrol (CFG). Metode penyematan dan augmentasi grafik digunakan untuk menghasilkan dan menyematkan contoh manipulasi ke dalam data pelatihan perangkat lunak IoT. Fitur CFG memungkinkan eksplorasi malware IoT melalui teori grafik dan pembelajaran mesin. Martins dkk. menganalisis pembuatan dan deteksi contoh manipulasi dalam skenario deteksi intrusi dan malware.

Untuk menghasilkan contoh manipulasi, algoritme pembelajaran mesin dikategorikan ke dalam simbolis seperti pohon keputusan, koneksionis seperti jaringan saraf, evolusioner seperti algoritme genetika, Bayesian seperti Naive Bayes, analogisator seperti k-nearest neighbours. Pertahanan manipulasi untuk malware dan deteksi intrusi diberikan sebagai pelatihan manipulasi, penyembunyian gradien, distilasi defensif, pemerasan fitur, blok transferabilitas, metode pertahanan gangguan universal, dan autoencoder MagNet.

Kecerdasan Buatan dalam Pembelajaran Mendalam Permainan Strategi

Tan dkk. membahas peta perhatian dalam tugas visi komputer. Prioritas geometris pada konteks spasial untuk sebuah piksel dimodelkan sebagai modul perhatian diri yang baru. Hal ini tidak memerlukan pengkodean posisi yang mahal secara komputasi dari peta perhatian berbasis konten yang dibuat dengan kueri dan kunci. Konsep pelatihan perhatian mandiri tidak hanya berlaku untuk tugas visi komputer tetapi juga tugas pemrosesan bahasa alami. Dalam tugas pengenalan gambar dikategorikan menjadi perhatian saluran dan perhatian spasial.

Sen dkk. melakukan penilaian kuantitatif mekanisme perhatian manusia versus komputasi dalam tugas klasifikasi teks. Mereka dikontraskan dalam pemilihan kata yang tumpang tindih, distribusi kategori leksikal, dan polaritas sentimen yang bergantung pada konteks. Mekanisme perhatian berguna untuk interpretasi tentang detail pemodelan seperti debugging model, pemilihan arsitektur dalam tugas pemrosesan bahasa alami (NLP) seperti pemodelan bahasa, terjemahan mesin, klasifikasi dokumen, dan menjawab pertanyaan. Mereka menciptakan skor perhatian yang dapat dijelaskan untuk prediksi model yang dapat dihubungkan dengan ukuran pentingnya fitur pada pengurangan dimensi.

Perhatian manusia diukur dari perspektif ukuran kesamaan perilaku, kesamaan leksikal (tata bahasa), dan polaritas sentimen ketergantungan konteks. Peta perhatian yang dihasilkan dibandingkan dengan jaringan saraf berulang (RNN) berbasis perhatian. RNN dua arah dengan mekanisme perhatian ditemukan serupa dengan perhatian manusia menurut ukuran perhatian manusia. Peta perhatian didefinisikan sebagai vektor dengan urutan kata yang dikaitkan dengan posisi dalam teks. Jaringan saraf dapat menghasilkan peta perhatian dengan menghitung distribusi probabilitas atau operasi bitwise pada rangkaian kata. Tugas prediksi NLP menjadi lebih sulit pada teks yang panjang karena skor akurasi dan kesamaan model menurun.

Lin dkk. membuat RankGAN untuk menghasilkan deskripsi bahasa alami dari kalimat yang ditulis manusia dan kalimat yang ditulis mesin. Diskriminator melakukan pemeringkatan relatif pada teks untuk membantu menciptakan generator yang lebih baik. Informasi peringkat relatif tersebut dapat memperoleh manfaat dari metode agregasi peringkat yang digunakan dalam pemerataan distribusi fitur pembelajaran adversarial. Peta perhatian dalam pembelajaran mendalam dapat dikontraskan dengan peta fitur seperti yang dijelaskan oleh Thaller et al. untuk menganalisis pola desain dalam masalah pengembangan perangkat lunak yang berulang.

Peta fitur adalah representasi perangkat lunak berdasarkan struktur mikro yang dapat dipahami manusia dan mesin. Ruang vektor di atas struktur mikro dalam peta fitur dapat

didefinisikan sebagai ruang fitur berdimensi tinggi untuk mendeteksi contoh pola desain dalam kode sumber dengan pembelajaran mesin. Dengan demikian, deskripsi pola berbasis pembelajaran mesin dapat digunakan untuk memecahkan masalah arsitektur berorientasi objek (OO) tingkat tinggi seputar pembuatan, struktur, atau perilaku kelas dan objek. Semantik dalam deskripsi pola berisi nama, maksud, motivasi yang berlaku pada struktur, partisipan, dan kolaborasi dalam kode sumber. Mereka digunakan untuk membuat keputusan desain dan alasan dokumentasi selama pengembangan produk perangkat lunak. Mengambil informasi yang dikodekan dengan deteksi pola desain (DPD) berguna dalam pengembangan kembali dan pemeliharaan produk perangkat lunak.

DPD menemukan struktur dan ketergantungan dalam kode sumber untuk menghasilkan grafik semantik abstrak (ASG) yang menyoroti algoritma dan bagian bergeraknya untuk meningkatkan kinerja sistem. Burnap dkk. mengembangkan peta fitur pengorganisasian mandiri (SOFM) untuk membedakan antara sampel perangkat lunak portabel yang dapat dieksekusi yang berbahaya dan tepercaya. Fitur pembelajaran mesin dibuat pada byte dan paket dalam jejak yang ditinggalkan oleh sistem komputer selama eksekusi melalui CPU, RAM, swap, dan jaringan. Berbeda dengan fitur yang berasal dari panggilan API, fitur eksekusi tersebut tidak dapat dikaburkan dengan mudah dalam serangan gaya APT.

SOFM menangkap lingkungan topografi dalam data yang dipisahkan oleh batas fuzzy antar kelas aktivitas mesin. SOFM mampu mengatasi dilema plastisitas-stabilitas untuk sistem pembelajaran yang perlu beradaptasi dengan lingkungan sambil mempertahankan kemanjuran fungsi stabil. Memanfaatkan lingkungan topografi sebagai kumpulan fitur fuzzy dalam algoritma klasifikasi mesin meningkatkan perilaku generalisasi algoritma pembelajaran pada sampel yang tidak terlihat dalam muatan berbahaya seperti malware polimorfik. SOFM juga dapat digunakan untuk visualisasi dan eksplorasi data di pusat operasi keamanan. Dotter dkk. mencoba mengaitkan masukan yang terganggu dengan metode serangan tertentu dalam upaya untuk mengekspos algoritma serangan, arsitektur model, dan hyperparameter yang digunakan dalam serangan melalui kerangka pembelajaran yang diawasi.

Indikator atribusi dunia maya diperoleh untuk taktik, teknik, dan prosedur (TTP) perdagangan dan malware yang meninggalkan sinyal dan tanda pengenal tertentu. Atribusi serangan tersebut dapat digunakan bersama dengan indikator perdagangan siber lainnya seperti niat dan infrastruktur. Di sini, teknik pembelajaran mendalam adversarial dalam desain pengklasifikasi atribusi dapat secara otomatis merekayasa balik rantai alat untuk atribusi adversarial di balik serangan siber seperti deepfake, pemalsuan multimedia, serangan pembelajaran mesin adversarial, dan serangan penipuan informasi. Atribusi serangan pada kumpulan data adversarial dinyatakan sebagai atribusi algoritma serangan, atribusi hyperparameter, atribusi model, dan atribusi norma.

Samek dkk. mengembangkan metode analisis sensitivitas untuk memvisualisasikan, menjelaskan, dan menafsirkan model pembelajaran mendalam untuk meningkatkan transparansi prediksi mereka. Interpretabilitas dan penjelasan aplikasi kecerdasan buatan yang dapat dipercaya merupakan disiplin baru dalam pembelajaran mesin yang menghitung

sensitivitas prediksi sehubungan dengan perubahan masukan. Model pembelajaran mendalam bertindak sebagai sistem kotak hitam secara default. Ada kebutuhan mendesak untuk memahami pembelajaran suatu model dan menjelaskan prediksi individualnya untuk memajukan model pembelajaran mesin di luar jaringan saraf.

Metode kecerdasan buatan yang dapat dijelaskan tersebut diperlukan dalam sistem pembelajaran mesin untuk verifikasi pengambilan keputusan sistem, peningkatan arsitektur sistem, transfer pengetahuan pembelajaran sistem ke pengguna manusia, dan kepatuhan keputusan algoritmik terhadap privasi. peraturan. Dengan demikian, hubungan antara kemampuan generalisasi, kekompakan, dan kemampuan menjelaskan representasi yang dipelajari dalam pembelajaran mendalam adversarial merupakan bidang penelitian yang aktif. Ancona dkk. menggunakan nilai-nilai Shapley dari teori permainan kooperatif untuk menetapkan skor relevansi dalam metode atribusi. Mereka mengukur "*relevansi*" atau "*kontribusi*" setiap fitur masukan dalam sampel masukan tertentu. Keluaran target dalam tugas klasifikasi dipilih menjadi prediksi dengan probabilitas keluaran tertinggi yang dikaitkan dengan bagian masukan yang paling relevan untuk prediksi. Skor relevansi juga berisi informasi untuk menilai masukan bukti yang mendukung atau menolak label kelas yang diprediksi.

Metode atribusi juga dapat terkena serangan manipulasi tanpa metrik kuantitatif yang dapat diandalkan berdasarkan kebenaran dasar untuk mengevaluasi penjelasannya. Di sini nilai-nilai Shapley bertindak sebagai properti penjelasan yang terbukti dengan sendirinya yang dirancang untuk jaminan teoretis yang lebih kuat atas keandalannya. Nilai Shapley dapat ditetapkan ke atribut sedemikian rupa sehingga aksioma tertentu yang diinginkan terpenuhi pada kelengkapan, simetri, linearitas, kontinuitas, dan invarian implementasi metode atribusi. Choras dkk. membahas kurangnya keadilan dan penjelasan dalam algoritma canggih untuk pembelajaran mesin dan kecerdasan buatan di beberapa domain aplikasi yang menggunakan kemampuan pembelajaran mendalam untuk menyelesaikan tugas deteksi atau prediksi. Di sini kerangka keamanan dalam pembelajaran mesin manipulasi dapat menimbulkan disinformasi untuk menyesatkan hasil pembelajaran mendalam.

Keadilan dalam kecerdasan buatan kemudian berkaitan dengan kerangka etika dan hukum seputar disinformasi yang dapat disebarkan secara jahat di masyarakat luas. Bias algoritmik yang diakibatkan oleh bias operator manusia yang memberikan data dengan representasi yang salah dan diskriminasi menyebabkan ketidakadilan dalam kecerdasan buatan. Di sini terdapat kebutuhan untuk membuat kumpulan data pelatihan tanpa sampel yang miring, contoh yang tercemar, dan fitur terbatas yang menyebabkan atribut bias sensitif dalam algoritme pelatihan dan selanjutnya disparitas ukuran sampel dalam algoritme klasifikasi. Jadi keadilan pembelajaran mesin harus didefinisikan berdasarkan pengertian ketidaksadaran, keadilan kelompok, dan keadilan kontrafaktual dalam formulasi matematis pembelajaran mendalam adversarial. Di sini kontrafaktual akibat pembelajaran mesin manipulasi dapat dimodelkan sebagai grafik sebab akibat yang menjelaskan prediksi pembelajaran mendalam yang diawasi.

Dalam konteks ini, prosedur pembelajaran mendalam adversarial teoretis permainan memberikan kerangka statistik untuk mengoptimalkan keseimbangan antara ukuran akurasi dan keadilan pada kinerja sistem pembelajaran mesin. Mereka dapat membuat pengklasifikasi yang adil dengan mengacu pada serangkaian masalah klasifikasi yang sensitif terhadap biaya yang memberikan pengklasifikasi acak dengan kesalahan empiris terendah dalam batasan optimasi yang diinginkan. Arrieta dkk. mensurvei literatur tentang AI yang dapat dijelaskan (XAI) dan memberikan taksonomi tentang kontribusi terkini dalam pembelajaran mendalam.

Hal ini mengarah pada konsep yang lebih luas tentang kecerdasan buatan yang bertanggung jawab seputar metodologi penerapan kecerdasan buatan dalam skala besar di organisasi dunia nyata dengan keadilan, penjelasan, dan akuntabilitas yang tertanam dalam kecerdasan buatan untuk setiap industri yang diatur di setiap sektor kegiatan ekonomi. Interpretabilitas sebagai pendorong desain dalam pembelajaran mesin mendukung ketidakberpihakan dalam pengambilan keputusan, memfasilitasi penyediaan ketahanan yang dapat dipelajari, dan bertindak sebagai jaminan atas kausalitas mendasar yang ada dalam penalaran model. Samek dkk. memberikan ulasan lain tentang XAI di jaringan saraf dalam.

Ribeiro dkk. mempelajari model lokal yang dapat ditafsirkan seputar prediksi pengklasifikasi. Ini menjelaskan prediksi individu sebagai solusi untuk masalah optimasi submodular. Kegunaan penjelasan tersebut divalidasi dalam eksperimen yang menilai kepercayaan pada kotak hitam pembelajaran mesin dengan memahami alasan di balik prediksi tersebut. Kerangka kerja penjelasan agnostik model yang dapat ditafsirkan secara lokal (LIME) disajikan untuk masalah “mempercayai prediksi” untuk mengambil tindakan berdasarkan prediksi tersebut, “mempercayai model” untuk berperilaku dengan cara yang wajar ketika diterapkan di dunia nyata. Representasi yang dapat diinterpretasikan untuk artefak tekstual dan visual dihasilkan sebagai tensor penjelasan untuk setiap contoh masukan sehingga kriteria interpretasi khusus domain dan tugas dapat diakomodasi.

LIME memiliki aplikasi dalam sistem rekomendasi untuk domain ucapan, video, dan medis untuk merancang sistem pembelajaran mesin human-in-the-loop. Hartl dkk. memperkenalkan ukuran sensitivitas fitur yang disebut skor ketahanan adversarial (ARS) untuk data aliran jaringan berurutan dalam sistem deteksi intrusi (IDS). Hal ini berguna sebagai ukuran pentingnya fitur yang digunakan dalam pembuatan sampel manipulasi untuk jaringan saraf berulang (RNN). ARS dapat digunakan bersamaan dengan akurasi untuk mengevaluasi sistem pembelajaran mesin yang sensitif terhadap keamanan. Ini meningkatkan metode penjelasan seperti plot ketergantungan parsial (PDP) untuk data sekuensial.

Mekanisme pertahanan yang diusulkan menggunakan ARS untuk menghilangkan fitur yang dapat dimanipulasi, mengurangi permukaan serangan, dan memperkuat IDS yang dihasilkan. Melis dkk. mengevaluasi kepercayaan terhadap pendeteksi malware Android saat pendeteksi tersebut bertransisi dari kinerja yang baik pada data benchmark menjadi penerapan di lingkungan operasi. Pendekatan berbasis gradien mengidentifikasi fitur lokal yang paling berpengaruh untuk meningkatkan akurasi tanpa kehilangan interpretasi

keputusan. Penjelasannya dapat memberikan wawasan tentang kerentanan model pembelajaran mesin blackbox yang digunakan untuk mendeteksi malware.

Demetrio dkk. memberikan atribusi fitur pada setiap keputusan yang dibuat untuk klasifikasi biner malware. Penjelasan tersebut kemudian digunakan untuk menghasilkan biner malware manipulasi yang lebih baik daripada algoritme serangan canggih terhadap algoritme pembelajaran mendalam yang menyediakan fungsi keputusan yang sangat non-linier. Kontribusi setiap fitur pada label titik data dihitung sehubungan dengan garis dasar yang menciptakan kebenaran dasar untuk pemodelan. Gangguan manipulasi kemudian meningkatkan kontribusi yang dihitung untuk keluaran pemodelan pada fitur-fitur yang dimodifikasi di garis dasar. Aksioma sensitivitas dibuat sebagai garis dasar untuk memandu algoritma pelatihan kesalahan propagasi balik melalui jaringan saraf.

Gradien terintegrasi digunakan untuk menjelaskan hasil klasifikasi. Namun, model yang dapat dijelaskan tetap rentan terhadap manipulasi yang merugikan. Marino dkk. menghasilkan penjelasan untuk kesalahan klasifikasi dalam sistem deteksi intrusi berbasis data. Penjelasan tersebut memberikan alasan di balik kesalahan klasifikasi dan mencocokkannya dengan pengetahuan ahli. Penjelasan ini berlaku untuk semua pengklasifikasi yang memiliki gradien. Mereka dapat digunakan dalam forensik digital dan penilaian kerentanan sistem pembelajaran mesin yang mendasarinya. XAI tersebut menggunakan visualisasi data dan deskripsi bahasa alami untuk menjelaskan alasan keputusan yang dibuat oleh sistem pembelajaran mesin.

Penalaran tersebut dapat dipahami oleh manusia dan digunakan untuk menyederhanakan proses penemuan pengetahuan dalam data. Ini juga dapat digunakan untuk menghasilkan diagnostik debugging pada sistem pembelajaran mesin. Penjelasan diasumsikan sebagai modifikasi minimum yang diperlukan untuk mengklasifikasikan sampel yang salah klasifikasi dengan benar. Manipulasi digunakan untuk memvisualisasikan fitur pembelajaran yang menyebabkan kesalahan klasifikasi. Kesalahan klasifikasi sering ditemukan terjadi antara sampel dengan karakteristik data yang bertentangan.

Liu dkk. menyelidiki interpretasi model untuk mendukung kerangka deteksi manipulasi yang menjelaskan prediksi dalam model pembelajaran mesin target. Pelatihan manipulasi kemudian digunakan untuk meningkatkan ketahanan detektor pada sampel manipulasi. Biaya manipulasi fitur diperkirakan untuk mengkategorikan tipe musuh. Kerangka kerja deteksi yang ada dikategorikan ke dalam metode rekayasa fitur yang rentan terhadap musuh adaptif, interaksi teoretis permainan antara detektor dan musuh yang spesifikasi pemodelannya bervariasi sesuai arsitektur pengklasifikasi pembelajaran mesin, dan pertahanan jaringan saraf dalam seperti pelatihan manipulasi, distilasi defensif, dan fitur tindihan.

Titik data yang salah klasifikasi diunggulkan dari sampel yang rentan terhadap penghindaran yang kemungkinan besar akan bergeser melintasi batas pengambilan keputusan. Serangan manipulasi dibangun berdasarkan arah gangguan berdasarkan interpretasi model dari contoh data masukan yang diklasifikasikan sebagai jinak atau berbahaya. Lundberg dkk. mengusulkan kerangka teori permainan yang disebut SHAP

(SHapley Additive exPlanations) untuk mempelajari trade-off antara akurasi dan interpretabilitas dalam hasil pembelajaran mendalam. SHAP menghitung ukuran pentingnya fitur aditif untuk setiap prediksi sebagai nilai regresi Shapley. Perkiraan pengambilan sampel dilakukan dalam perhitungan nilai Shapley.

Metode estimasi nilai Shapley ditambah dengan metode atribusi fitur yang memenuhi properti yang diinginkan pada penjelasan seperti keakuratan lokal yang mencocokkan model penjelasan dengan model asli, ketidakhadiran untuk mencegah fitur yang hilang berdampak, dan konsistensi pada atribusi masukan sehubungan dengan perubahan dalam model. keadaan karena masukan lainnya. Kemudian teori permainan kooperatif digunakan untuk membuktikan secara matematis tidak melanggar persyaratan akurasi dan interpretabilitas di mana nilai Shapley bertindak sebagai fungsi ekspektasi bersyarat dari pentingnya fitur dalam model asli.

Fungsi ekspektasi bersyarat didekati dengan metode khusus model seperti nilai pengambilan sampel Shapley, Kernel SHAP, Max SHAP, dan Deep SHAP. Perkiraan model-agnostik diperoleh dari metode pengaruh masukan kuantitatif yang merupakan perkiraan pengambilan sampel versi permutasi persamaan nilai Shapley klasik. Estimasi gabungan nilai SHAP dengan regresi memberikan kompleksitas/efisiensi sampel yang lebih baik daripada penggunaan langsung persamaan Shapley klasik. Dengan demikian, penjelasan teoretis permainan tentang pembelajaran mendalam adversarial memberikan jalan untuk menciptakan kelas model penjelasan baru.

Beyazit dkk. mengusulkan representasi yang dapat ditafsirkan yang dipelajari melalui model generatif mendalam dengan mengekstraksi fitur marginal independen serta keterikatan kausalitas dalam data pelatihan. Pengatur pelatihan kemudian menghukum ketidaksepakatan antara interaksi fitur yang diekstraksi dan struktur ketergantungan tertentu dalam data pelatihan. Pengatur menerapkan batasan struktural pada ruang laten interaksi fitur untuk memberikan kinerja generalisasi yang lebih baik daripada yang canggih. Interaksi fitur menggunakan jaringan Bayesian untuk menghitung parameter kemungkinan maksimum yang mengukur ketidaksepakatan antara ruang laten dan ruang pelatihan.

Maksimalisasi informasi timbal balik di InfoGAN mengekstrak fitur-fitur yang bermakna secara visual dan manipulasinya. Model struktur ketergantungan adalah hubungan antara fitur data yang diamati dan yang menonjol. Struktur ketergantungan bertindak sebagai kendala pembelajaran tambahan pada pelatihan InfoGAN. Fungsi kemungkinan untuk generator mengukur probabilitas data pelatihan berdasarkan model generator data terbaik. Umpan maju InfoGAN dipandang sebagai pemetaan dari ruang variabel yang diamati hingga suatu keputusan. Fungsi tujuan untuk distribusi penghasil data menghasilkan instance data yang sesuai dengan mean squared error (MSE) untuk model Gaussian linier.

Fungsi kerugian struktural bertindak sebagai pengatur pelatihan InfoGAN. Struktur grafik yang optimal dapat dirancang dalam ruang laten untuk mengeksplorasi fitur-fitur yang menonjol. Representasi yang dapat ditafsirkan seperti itu dapat digunakan dalam pembelajaran transfer, pembelajaran zero-shot, dan pembelajaran penguatan. Molnar mensurvei metode model-agnostik untuk menafsirkan model kotak hitam dalam

pembelajaran mendalam. Ini termasuk pentingnya fitur, akumulasi efek lokal, nilai-nilai Shapley, dan LIME. Hasil dari metode interpretasi meliputi statistik ringkasan fitur, visualisasi ringkasan fitur, bobot yang dipelajari, penjelasan kontrafaktual, dan kotak putih.

Biasanya metode interpretasi untuk jaringan neural dalam bersifat spesifik model dan terbatas pada kelas model tertentu. Metode model-agnostik untuk interpretasi teknik pembelajaran mesin secara umum mencakup plot ketergantungan parsial, akumulasi plot efek lokal, interaksi fitur (statistik H), dekomposisi fungsional, permutasi pentingnya fitur, dan model pengganti global. Interpretasi jaringan saraf diekspresikan dalam bentuk fitur yang dipelajari, atribusi piksel (peta arti-penting), contoh yang berpengaruh, penjelasan kontrafaktual, dan contoh manipulasi.

Bitton dkk. melakukan analisis ancaman terhadap sistem produksi pembelajaran mesin. Model ancaman menghitung aset pembelajaran mesin, potensi musuh, tujuan manipulasi, tujuan pembelajaran, dan skenario serangan dalam sistem pembelajaran. Sistem penilaian dirancang untuk berbagai serangan manipulasi. Ini menggunakan proses hierarki analitik (AHP) untuk menentukan peringkat atribut serangan para pakar keamanan siber. Kemudian kerangka pembuatan grafik serangan yang disebut MulVAL dikembangkan sebagai grafik serangan logis untuk menggabungkan efek serangan siber dalam sistem produksi pembelajaran mesin untuk beberapa kasus penggunaan dalam keamanan siber, deteksi penipuan, perdagangan keuangan, pemasaran yang dipersonalisasi, optimalisasi sumber daya, layanan kesehatan, dan kendaraan otonom.

Keputusan penting yang menyesatkan yang dibuat terkait kasus penggunaan ini mempunyai dampak signifikan secara statistik terhadap perencanaan darurat, kelangsungan bisnis, aliran pendapatan, dan bahkan kehidupan manusia. Selain bug dalam sistem pembelajaran mesin, bug tersebut juga harus menghadapi kerentanan logis dalam algoritma ML yang mendasarinya. Musuh dapat mengeksploitasi bug dan kerentanan tersebut sebagai skenario serangan untuk pembelajaran mesin manipulasi. Oleh karena itu, sistem pembelajaran mesin harus dilengkapi dengan alat taktis dan strategis untuk menganalisis, mendeteksi, melindungi, dan merespons serangan siber. Alat-alat tersebut telah dikembangkan sebagai kerangka kerja dan perpustakaan seperti perpustakaan contoh manipulasi CleverHans, Kotak Alat Ketahanan Adversarial, kotak alat Foolbox, perpustakaan SecML, dan MLsploit. Saat ini mereka menerapkan algoritma untuk menghasilkan dan membedakan contoh-contoh manipulasi.

Namun hal ini harus diperluas untuk mengukur risiko sistem pembelajaran mesin, melakukan pemodelan ancaman keamanan siber dalam penerapan tertentu dan lingkungan target untuk algoritme pembelajaran mesin, dan mengukur berbagai atribut teknik serangan seperti model penyerang, dampak serangan, dan kinerja serangan. Di sini karakteristik pipeline produksi pembelajaran mesin dapat dinyatakan sebagai elemen data, penerapan, pengiriman, dan orkestrasi untuk membuat ontologi analisis ancaman untuk aset, kerentanan, penyerang, kemampuan, dampak, ancaman, dan teknik serangan. Kerangka kerja MulVAL menganalisis grafik serangan logis untuk secara otomatis mengekstrak informasi dari database kerentanan formal dan alat pemindaian jaringan. Ini menghitung semua

kemungkinan jalur serangan dalam waktu polinomial pada skenario serangan baru dan yang sedang berkembang.

MulVAL berguna dalam desain algoritma penilaian risiko dan perencanaan penanggulangan yang divalidasi dengan aturan interaksi eksplisit dan predikat untuk pemodelan serangan dalam bahasa pemrograman Datalog. Elitzur dkk. menganalisis intelijen ancaman siber (CTI) pada serangan sebelumnya untuk rekonstruksi serangan dalam alat pada pola serangan yang tidak teramati yang dapat menambah korelasi peringatan dan visualisasi data untuk analis keamanan siber yang mempelajari hipotesis serangan dalam forensik digital pembelajaran mesin manipulasi dengan rantai pembunuhan siber. Generator Hipotesis Serangan (AHG) membuat grafik pengetahuan tentang intelijen ancaman untuk menghasilkan hipotesis serangan dalam manajemen informasi dan peristiwa keamanan (SIEM). Di sini CTI dikategorikan menjadi intelijen ancaman strategis, intelijen ancaman operasional, intelijen ancaman taktis, dan intelijen ancaman teknis.

Mereka digunakan untuk membangun grafik pengetahuan untuk mendukung penalaran manipulasi dengan fitur penambangan grafik pada kesamaan topologi, korelasi, dan pola frekuensi. Mengeksekusi Aturan Web Semantik Bahasa pada grafik pengetahuan dapat digunakan dalam analitik berbasis data dari aturan inferensi deduktif berbasis logika. Prediksi tautan dan pemfilteran kolaboratif dalam grafik pengetahuan dapat meningkatkan pembuatan hipotesis serangan dengan skenario serangan yang mungkin terjadi. Matern dkk. membuat artefak visual yang dapat digunakan dalam alat forensik statistik untuk mengungkap manipulasi di Deepfakes. Tujuan manipulasi dalam pembuatan video otomatis adalah untuk menciptakan manipulasi jahat untuk menyampaikan pesan semantik dalam video yang awalnya tidak dimaksudkan dalam materi pelatihan. Di sini forensik gambar mencari artefak gambar fisik atau statistik untuk membentuk sidik jari statistik, memvalidasi noise prior, dan mempelajari jejak manipulasi spesifik pada sisa gambar untuk mendeteksi manipulasi.

Artefak visual yang diusulkan dikategorikan ke dalam masalah visi komputer seperti konsistensi global, estimasi iluminasi, dan estimasi geometri. Kamath dkk. menjawab pertanyaan teoretis dalam pembelajaran statistik tentang bagaimana suatu distribusi dapat didekati dengan sampelnya. Ukuran kerugian yang halus diusulkan untuk perkiraan distribusi. Untuk aplikasi kompresi dan investasi, kerugian yang relevan adalah divergensi Kullback-Leibler (KL). Untuk klasifikasinya adalah L1, L2, Hellinger, chi-squared, softmax loss. Untuk pembelajaran adversarial, kasus terburuk yang paling kecil untuk penaksir optimal teori permainan untuk pembelajaran mendalam adversarial adalah kerugian minmax. Untuk pembelajaran online di lingkungan dengan sumber daya terbatas, kerugiannya adalah kerugian kumulatif minimal berdasarkan statistik dan teori informasi untuk meminimalkan kerugian dibandingkan perkiraan berturut-turut.

Katzir dkk. mengukur ketahanan sistem pembelajaran mesin dengan metode formal yang berlaku untuk sistem fusi multisensor. Skor ketahanan model (MRB) diusulkan untuk mengevaluasi ketahanan manipulasi guna mengontrol trade-off ketahanan vs akurasi dalam klasifikasi malware dinamis. Pemilihan fitur yang sadar akan musuh bertujuan untuk

menemukan subkumpulan fitur yang anggaran musuh tidak mencukupi untuk menciptakan manipulasi, generalisasi pengklasifikasi dimaksimalkan, dan dimensi pelatihan diminimalkan. Kemudian MRB digunakan sebagai kriteria pemilihan fitur resilient. Evaluasi eksperimental kemudian dilakukan terhadap biaya manipulasi fitur bagi musuh untuk menargetkan fitur tangguh dan tidak tangguh dari pengklasifikasi tangguh musuh dalam sistem pertahanan siber dengan fusi multisensor.

Sadeghi dkk. mensurvei titik temu antara kecerdasan komputasi dan pembelajaran mesin pada kendaraan otonom, robot bantu, dan sistem biometrik. Di sini kesalahan klasifikasi akibat serangan manipulasi mengakibatkan keputusan yang salah dan operasi yang tidak dapat diandalkan. Sistem pembelajaran mesin manipulasi dapat dikategorikan secara terperinci berdasarkan kumpulan data masukan, arsitektur ML, spesifikasi musuh, metodologi pembangkitan serangan, dan strategi pertahanan. Cho dkk. mengusulkan metrik keamanan dan ketergantungan sebagai metrik utama untuk membangun sistem pembelajaran mesin yang dapat dipercaya dalam lingkungan multi-domain yang mencakup perangkat keras, perangkat lunak, jaringan, faktor manusia, dan lingkungan fisik.

Kerangka metrik kepercayaan mendukung kerangka kerja berbasis ontologi untuk kepercayaan, ketahanan, dan ketangkasan. Hal ini dapat digunakan dalam penilaian kerentanan, tim merah komputasi, dan pengukuran sistem yang dapat dipercaya. Metrik (atau pengukuran) yang dapat dipercaya dapat digunakan dalam validasi pembelajaran mesin manipulasi untuk objektivitas berdasarkan kepastian, efisiensi berdasarkan kuantifikasi, dan kontrol berdasarkan umpan balik. Mereka bertindak sebagai atribut data tentang kualitas sistem seperti kegunaan, pengelolaan, fungsionalitas, kinerja, ketergantungan, kemampuan beradaptasi, keamanan, dan biaya. Mereka dapat memasukkan persyaratan keamanan seperti ketersediaan, integritas, kerahasiaan, keandalan, ketersediaan, integritas, keselamatan, dan pemeliharaan. Teknik asal data juga dapat disertakan dengan metrik yang dapat dipercaya untuk mengevaluasi pembagian informasi yang dapat dipercaya dalam sistem sosio-teknis dan penginderaan siber.

Definisi asal data dapat digunakan untuk membuktikan pernyataan kualitas data. Mereka menghubungkan keandalan dan reproduktifitas aplikasi analisis data dengan asal dan kepemilikan data, validasi pemodelan, dan pembenaran atas hasil yang tidak diharapkan. Kamu dkk. membuat metode pembelajaran jarak jauh khusus yang cocok untuk optima lokal berbeda pada algoritma manipulasi yang dirancang untuk properti statistik dalam distribusi data sebenarnya. ISMETS (Subruang METric Khusus Instance) yang diusulkan mencakup seluruh ruang metrik untuk pembelajaran jarak jauh secara generatif. Ia mempelajari subruang metrik untuk setiap contoh dengan menyimpulkan ekspektasi distribusi dalam inferensi variasional atas basis metrik menurut paradigma Bayesian untuk induksi dan transduksi.

Subruang metrik berguna untuk memahami kemampuan interpretasi dan ketahanan hasil pembelajaran mendalam manipulasi dengan variabel alokasi laten yang menggabungkan informasi sampingan. Teknik pemrograman paralel dan metode perkiraan numerik juga dapat dimasukkan ke dalam kerangka pembelajaran metrik untuk

memperluasnya ke kumpulan data berdimensi tinggi. Algoritme pembelajaran metrik jarak jauh seperti itu selanjutnya dapat digunakan untuk membatasi biaya pengoptimalan dalam fungsi pembayaran adversarial untuk menyelesaikan masalah pengoptimalan multi-tujuan, terbatas, berskala besar, dan tidak pasti untuk pembelajaran mendalam adversarial. Sugiyama dkk. membahas pentingnya menghitung pendekatan divergensi antara jenis distribusi probabilitas dari sampelnya dalam pembelajaran mesin, teori informasi, dan statistik.

Estimator divergensi memiliki aplikasi penambangan data pada distribusi analisis data seperti deteksi titik perubahan, estimasi keseimbangan kelas, pemilihan dan ekstraksi fitur, pengelompokan, pencocokan objek, analisis komponen independen, dan estimasi arah sebab akibat. Divergensi diperkirakan dengan cara komputasi yang efisien tanpa memperkirakan distribusi probabilitas yang mendasarinya. Meskipun divergensi Kullback-Leibler adalah pendekatan divergensi yang paling populer, pendekatan lain seperti divergensi Pearson dan ukuran jarak L2 juga berguna dalam pembelajaran mesin karena sifat stabilitas dan ketahanannya. Ukuran divergensi dapat disebut sebagai jarak untuk pembelajaran metrik jika ukuran tersebut memenuhi sifat matematika non-negatif, non-degenerasi, simetri, dan pertidaksamaan segitiga.

Dalam pembelajaran generatif manipulasi, kami ingin mencapai konvergensi terhadap perkiraan distribusi target dengan pembelajaran mendalam. Oleh karena itu ukuran jarak statistik dan metode distribusi relatif merupakan ukuran yang cocok untuk mengevaluasi kinerja generalisasi dari distribusi target. Tzeng dkk. menerapkan metode pembelajaran manipulasi pada adaptasi domain untuk memahami pergeseran domain karena kumpulan data/bias algoritmik. Oleh karena itu, pembelajaran mendalam yang bermusuhan dapat digunakan untuk menghasilkan sampel yang kompleks di berbagai domain. Performa generalisasi pembelajaran mendalam adversarial dapat ditingkatkan dengan meminimalkan perbedaan antara domain pelatihan, pengujian, dan validasi dengan kerugian adversarial yang dirancang sesuai. Jadi adaptasi adversarial menggeneralisasi pendekatan-pendekatan sebelumnya terhadap adaptasi domain.

Adaptasi manipulasi yang disebut Adaptasi Domain Diskriminatif Adversarial (ADDA) diusulkan oleh Tzeng dkk. ADDA dapat digunakan untuk mewakili domain sumber dan target dalam ruang fitur umum. Ini juga dapat digunakan untuk merekonstruksi domain target dari representasi sumber. Di sini adaptasi adversarial yang digeneralisasikan meminimalkan perkiraan perbedaan domain dalam tujuan adversarial untuk algoritma pembelajaran. Kami juga dapat mengukur perbedaan informasi antara representasi minimal data pelatihan dan penyematan fitur data adversarial dengan fungsi biaya adversarial berbasis pembelajaran metrik yang mendalam untuk adaptasi domain. Kami juga dapat menerapkan distribusi sebelumnya pada faktor laten untuk pembuatan data yang koheren dalam pembelajaran mendalam yang diawasi.

Hayes dkk. menerapkan pelatihan manipulasi untuk sintesis gambar dalam algoritma steganografi yang dinyatakan sebagai tugas pembelajaran diskriminatif untuk membangun steganalyzer yang kuat. Steganografi berkaitan dengan penyembunyian informasi dengan

menyematkannya dalam media non-rahasia. Steganografi dan kriptografi menyediakan metode menjaga privasi untuk komunikasi rahasia. Pesan steganografi dienkripsi dengan metode kriptografi sebelum ditanamkan pada media non-rahasia seperti pesan sampul dengan teks dan gambar. Pesan yang disematkan secara statistik tidak berbeda dengan string acak.

Pesan steganografi kemudian diterjemahkan untuk mengungkapkan ciphertext pesan tersebut. Teks sandi kemudian didekripsi dengan kunci kriptografi. Algoritma steganografi kemudian berupaya meminimalkan gangguan pada media penyematan. Tugas pembelajaran diskriminatif yang diusulkan menyematkan pesan rahasia dalam pesan sampul dengan algoritma untuk menghasilkan gambar steganografi. Musuh mempelajari kelemahan dalam algoritma penyematan untuk membedakan gambar sampul dari gambar steganografi. Oleh karena itu pelatihan adversarial dalam teknik steganalisis digunakan untuk memodelkan distribusi cover dengan benar. Modesitt dkk. menggabungkan kriptografi dengan pembelajaran mendalam yang bermusuhan untuk aplikasi dalam kriptanalisis dan enkripsi. Jaringan saraf dalam melakukan enkripsi simetris dalam lingkungan yang bermusuhan. Mereka memainkan permainan kriptografi dengan musuh untuk mendeteksi komunikasi yang tidak aman secara kriptografis berdasarkan ketidakmampuan membedakan ciphertext.

Steganografi Neural kemudian disediakan untuk membuat algoritma steganografi dengan adanya jaringan musuh. Krause dkk. mempelajari pengklasifikasi diskriminatif probabilistik yang disebut Regularized Information Maximization (RIM) dari kumpulan data yang tidak berlabel dan berlabel sebagian. RIM memiliki fungsi tujuan teori informasi untuk menyeimbangkan pemisahan kelas, keseimbangan kelas, dan kompleksitas pengklasifikasi dalam fungsi kemungkinan bersyarat kelas yang berbeda. RIM juga dapat diartikan sebagai algoritma clustering yang diskriminatif untuk merepresentasikan batasan antar kategori. Algoritme pengelompokan diskriminatif yang ada seperti partisi grafik spektral dan pengelompokan margin maksimum bukanlah model probabilistik seperti RIM. RIM memaksimalkan informasi timbal balik antara distribusi empiris pada input dan distribusi label terinduksi yang diatur dengan penalti kompleksitas. Istilah regularisasi membatasi batasan keputusan yang kompleks untuk menghasilkan solusi pengelompokan yang masuk akal.

Entropi relatif digunakan untuk mengakomodasi keyakinan sebelumnya tentang distribusi label dalam masalah klasifikasi kelas jamak untuk pembelajaran semi-supervisi sebagai istilah regularisasi lintas entropi. RIM mengarah pada prosedur optimasi yang efisien dan terukur untuk pemilihan model otomatis yang menentukan jumlah cluster. Hasil pengelompokan dibandingkan dengan label kebenaran dasar pada kategori kumpulan data dengan indeks Rand yang disesuaikan (ARI) yang membandingkan kluster inferensi statistik dengan label kebenaran dasar. Metode semi-supervisi yang dihasilkan terbukti secara signifikan meningkatkan kinerja klasifikasi garis dasar yang diawasi ketika jumlah contoh yang diberi label sedikit.

BAB 6
SERANGAN DUNIA FISIK TERHADAP GAMBAR DAN TEKS

Selama beberapa dekade terakhir, jaringan saraf dalam (DNN) telah menunjukkan keberhasilan besar dalam berbagai aplikasi, termasuk klasifikasi gambar dalam domain computer vision (CV) dan pengenalan teks dalam pemrosesan bahasa alami (NLP).) bidang. Namun, penelitian terbaru menunjukkan bahwa DNN sangat rapuh terhadap contoh-contoh yang merugikan terutama dalam domain gambar. Misalnya, Goodfellow dkk. menunjukkan bahwa menambahkan hampir nol noise pada gambar panda dapat menyesatkan GoogLeNet ke label yang salah (owa) dengan keyakinan tinggi (99,3%). Fenomena ini menimbulkan kekhawatiran besar mengenai implementasi keamanan DNN dan menarik banyak perhatian di komunitas CV sejak tahun 2014. Dalam literatur, banyak pendekatan telah diusulkan untuk menghasilkan contoh konflik untuk menyerang DNN (alias, cabang serangan) dan merancang mekanisme yang sesuai untuk mempertahankannya. potensi serangan (alias, cabang pertahanan). Dalam bab ini, kami fokus pada arah serangan untuk membuat contoh manipulasi berkualitas tinggi di domain CV dan domain NLP.

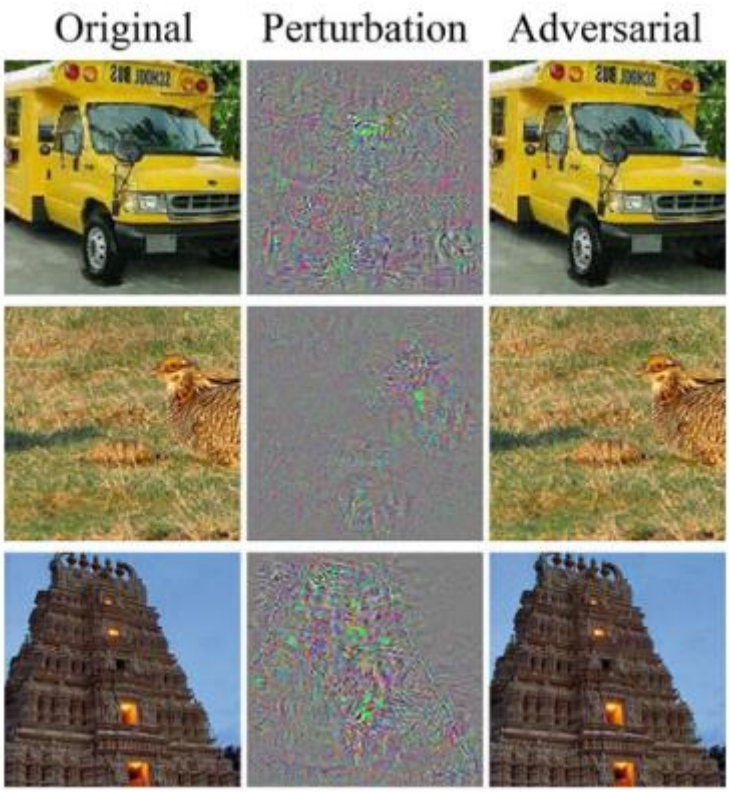
6.1 SERANGAN MANIPULASI TERHADAP GAMBAR

Serangan manipulasi adalah strategi efektif untuk menyelidiki properti DNN dan meningkatkan keamanan dan integritas aplikasinya. Serangan konflik pada gambar bertujuan untuk menghasilkan contoh manipulasi dengan menambahkan gangguan piksel halus pada gambar bersih sehingga model pembelajaran mendalam yang terlatih membuat prediksi yang salah. Dalam klasifikasi gambar, Szegedy et al. pertama kali mengungkapkan kelemahan yang menarik yaitu pemetaan input-output DNN sangat terputus-putus gangguan yang tidak terlihat oleh manusia sudah cukup untuk menyebabkan jaringan saraf membuat kesalahan klasifikasi (Gambar. 6.1). Lebih buruk lagi, gangguan yang sama pada gambar dapat menipu beberapa DNN meskipun mereka memiliki arsitektur jaringan yang berbeda. Hasil ini menyiratkan bahwa DNN saat ini sangat tidak stabil terhadap serangan musuh dan menarik minat besar dalam komunitas CV.

Berdasarkan, sejumlah pendekatan untuk serangan manipulasi gambar telah diusulkan, seperti serangan berbasis gradien, serangan berbasis skor, serangan berbasis keputusan, dan serangan berbasis transformasi. Sebagian besar strategi serangan ini menghitung gangguan untuk setiap gambar dengan menggunakan kumpulan data yang ada. Dibandingkan dengan serangan gambar tunggal, membuat gangguan universal untuk sekelompok gambar yang termasuk dalam kelas yang sama lebih menantang. Selain itu, sebagian besar mekanisme serangan yang ada dievaluasi pada kumpulan data publik, bukan pada lingkungan fisik, yang pengaturannya lebih kompleks. Dalam karakter ini, kami memperkenalkan modul serangan agnostik gambar baru untuk menghasilkan gangguan alami pada serangan rambu lalu lintas. Modul serangan ini dapat menghasilkan gangguan universal

untuk sekelompok rambu jalan, yang layak untuk diterapkan di dunia nyata. Hasil empiris pada kumpulan data publik dan gambaran dunia fisik menunjukkan bahwa metode ini mengungguli data dasar dalam hal tingkat keberhasilan serangan dan biaya gangguan. Dengan menggunakan modul perhatian lembut, ini menghasilkan gangguan yang lebih alami, yang terlihat seperti bayangan pohon oleh pengemudi manusia.

Pada bagian ini, kami meninjau empat jenis metode serangan manipulasi pada gambar, yaitu serangan berbasis gradien, serangan berbasis skor, serangan berbasis keputusan, dan transformasi- serangan berbasis.



Gambar 6.1 Contoh manipulasi yang berhasil dari untuk menyesatkan AlexNet. Gangguan tersebut hampir tidak terlihat oleh sistem penglihatan manusia, namun AlexNet memperkirakan contoh manipulasi tersebut adalah *"burung unta, struthio, camelus"* dari atas ke bawah.

Serangan Berbasis Gradien

Serangan berbasis gradien mencari arah gangguan yang paling sensitif untuk data masukan sesuai dengan fungsi gradien kerugian. Teman baik dkk. mengusulkan metode tanda gradien cepat (FGSM) yang terkenal, yang menentukan arah gangguan (peningkatan atau penurunan) untuk setiap piksel dengan memanfaatkan gradien fungsi kerugian. Mereka berpendapat bahwa kerentanan jaringan saraf disebabkan oleh sifat liniernya, bukan non-linier atau overfitting. Untuk mencapai efisiensi, FGSM dirancang untuk mempelajari gangguan melalui satu langkah gradien. Meskipun prosedur ini mempercepat pelatihan adversarial, namun sering kali gagal menemukan gangguan minimal dan menghasilkan biaya gangguan yang tinggi.

Kurakin dkk. menyempurnakan FGSM dengan mengulangi langkah gradien berkali-kali dengan ukuran langkah yang lebih kecil di setiap iterasi. FGSM berulang (I-FGSM) ini menyesatkan pengklasifikasi pada tingkat yang lebih tinggi dengan gangguan yang relatif lebih kecil. Kurakin dkk. juga menunjukkan bahwa usulan I-FGSM dapat menyesatkan pengklasifikasi target bahkan untuk sistem dunia fisik. Secara khusus, mereka mencetak contoh manipulasi yang dihasilkan di atas kertas dan mengambil fotonya dengan kamera ponsel. Hasil yang dilaporkan menguraikan bahwa sebagian besar foto-foto ini salah diklasifikasikan oleh pengklasifikasi ImageNet Inception. Metode DeepFool selanjutnya mengurangi kekuatan gangguan dengan mencari jarak secara berulang antara input bersih ke hyperplane klasifikasi terdekat. Namun, strategi optimasi serakah di I-FGSM dan DeepFool dengan mudah mengarah ke optimal lokal.

Dong dkk. merancang momentum I-FGSM (MI-FGSM), yang menggunakan vektor kecepatan untuk mengingat semua gradien sebelumnya selama iterasi untuk menghindari maksimum lokal yang buruk. Selain serangan kotak putih, Dong dkk. juga mengeksplorasi serangan blackbox dengan meningkatkan kemampuan transfer contoh-contoh manipulasi. Untuk meningkatkan kemampuan transfer, mereka mempelajari metode momentum iteratif untuk menyerang sekumpulan model, bukan hanya satu model. Landasan teoretisnya adalah jika contoh manipulasi yang dihasilkan dapat menipu semua model ansambel, kemungkinan besar serangan tersebut akan berhasil pada model yang tidak diketahui, karena kemampuan transfer adalah fakta bahwa model pembelajaran mesin yang berbeda mempelajari batasan keputusan yang serupa di sekitar titik data.

Baru-baru ini, Xiang dkk. menyematkan FGSM ke dalam skema serangan kotak abu-abu, di mana struktur jaringan korban tidak dapat diakses tetapi dapat diperoleh melalui serangan saluran samping (SCA). Secara khusus, SCA adalah teknik yang memperoleh pengetahuan internal melalui informasi saluran samping perangkat keras, seperti konsumsi waktu/daya dan radiasi elektromagnetik. Meskipun SCA tidak dapat mengungkapkan secara pasti bobot parameter atau fungsi kerugian, SCA dapat memperoleh struktur dasar jaringan. Oleh karena itu, serangan ini lebih praktis dibandingkan serangan white-box, karena struktur jaringan biasanya tidak diketahui, namun lebih unggul dibandingkan model blackbox yang tidak memiliki informasi.

Serangan Berbasis Skor

Serangan berbasis skor mengandalkan skor keluaran (misalnya, probabilitas yang diprediksi) alih-alih informasi gradien dalam membangun gangguan manipulasi, tanpa akses arsitektur model dan bobot model. Narodytska dkk. menerapkan skor kepercayaan untuk memandu metode pencarian lokal serakah, yang menemukan beberapa piksel (bahkan piksel tunggal) yang paling membantu dalam menghasilkan gangguan. Ini mengadopsi kriteria "kesalahan klasifikasi k teratas", yang berarti prosedur pencarian akan berhenti sampai label yang benar dikeluarkan dari skor k teratas. Salah satu kelemahan serangan piksel tunggal adalah piksel yang terganggu mungkin berada di luar jangkauan yang diharapkan.

Hayes dan Danezis melatih jaringan saraf penyerang untuk mempelajari gangguan, yang kemudian digunakan untuk menyerang jaringan target blackbox lainnya. Model

penyerang dilatih untuk meminimalkan perbedaan antara gambar masukan asli dan gambar keluaran musuh, di mana gambar keluaran dapat menyesatkan model target. Untuk mencapai hal ini, mereka mendefinisikan fungsi kerugian dengan menggabungkan skor kepercayaan keluaran dari kedua jaringan, yaitu kerugian rekonstruksi dan kerugian kesalahan klasifikasi. Kerugian rekonstruksi mengukur jarak antara masukan dan keluaran model penyerang untuk memastikan keluaran konflik terlihat serupa dengan masukan bersih. Kerugian kesalahan klasifikasi ditentukan berdasarkan jenis serangan (bertarget atau tidak bertarget) untuk membuat tingkat keberhasilan serangan tinggi.

Ilyas dkk. mempertimbangkan tiga skenario yang lebih realistis daripada pengaturan kotak hitam pada umumnya, termasuk (1) pengaturan terbatas kueri, (2) pengaturan informasi parsial, dan (3) pengaturan hanya label. Secara khusus, pengaturan terbatas kueri berarti penyerang memiliki jumlah kueri yang terbatas pada pengklasifikasi, pengaturan informasi parsial menunjukkan bahwa musuh hanya mengetahui probabilitas k teratas, dan pengaturan hanya label menunjukkan bahwa penyerang hanya memiliki akses ke k teratas. - k memberi label tetapi tidak mengetahui probabilitasnya. Untuk pengaturan kueri terbatas, penulis menggunakan strategi evolusi alami (NES) untuk memperkirakan gradien dan menghasilkan contoh manipulasi. Untuk mengatasi setting query-limited, penulis memulai dari instance kelas target dan bukan dari input asli, sehingga kelas top-k akan muncul di hasil prediksi. Untuk pengaturan label saja, mereka selanjutnya mendefinisikan perkiraan Monte Carlo untuk memperkirakan skor proksi probabilitas softmax.

Berdasarkan, Zhao dkk. mengusulkan algoritma penurunan gradien alami orde nol (ZO-NGD) untuk melakukan serangan manipulasi. Secara khusus, ini mengalikan gradien alami dengan matriks informasi Fisher (FIM) untuk mengoptimalkan model probabilistik. Kemudian menggabungkan FIM dengan penurunan gradien alami (NGD) orde kedua untuk mencapai efisiensi kueri yang tinggi.

Serangan Berbasis Keputusan

Serangan berbasis keputusan hanya memerlukan keputusan klasifikasi model (yaitu, label kelas 1 teratas) dan membebaskan kebutuhan gradien model atau skor keluarannya. Salah satu pekerjaan yang umum adalah serangan batas, yang dimulai dengan titik manipulasi, yaitu gambar yang dipilih dari kelas target. Kemudian mengurangi kebisingan dengan menerapkan jalan acak di sepanjang batas keputusan sambil tetap bersikap bermusuhan. Metode ini menambahkan gangguan minimal (dalam hal jarak L2) dibandingkan dengan metode berbasis gradien dan hampir tidak memerlukan hyperparameter untuk melakukan penyesuaian. Namun, diperlukan lebih banyak iterasi untuk menghasilkan contoh akhir yang berlawanan karena konvergensi yang lambat.

Berbeda dengan serangan batas yang meminimalkan gangguan dalam hal norma L2, Schott et al. mengusulkan serangan berbasis keputusan baru, yaitu serangan pointwise, yang mengurangi kebisingan dengan meminimalkan norma L0. Ini pertama-tama menginisialisasi titik awal dengan derau garam-merica atau derau Gaussian hingga gambar mengalami kesalahan klasifikasi. Kemudian berulang kali mengatur ulang setiap piksel yang terganggu

untuk membersihkan gambar sambil memastikan gambar yang berisik tetap bermusuhan. Prosedur ini berlangsung terus hingga tidak ada lagi piksel yang dapat direset.

Chen dan Jordan mengembangkan serangan batas dan mengusulkan estimasi arah gradien yang tidak bias pada batas keputusan menggunakan pencarian biner. Mereka menganalisis kesalahan estimasi ketika sampel tidak tepat berada pada batas dan menamai metode mereka sebagai Boundary Attack ++. Dibandingkan dengan serangan batas, Boundary Attack ++ tidak hanya mengurangi jumlah kueri model tetapi juga mampu beralih antara jarak L2 dan L dengan merancang dua operator klip.

Dalam, Chen dkk. menggunakan informasi biner dari batas keputusan untuk memperkirakan arah gradien dan menyajikan HopSkipJumpAtack (HSJA) berbasis keputusan. Metode ini dirancang untuk serangan yang ditargetkan atau tidak ditargetkan dengan meminimalkan jarak L2 atau L. Secara khusus, HSJA merupakan algoritma iteratif, dimana setiap iterasi berisi tiga langkah: estimasi arah gradien, pencarian ukuran langkah geometris, dan metode biner untuk pencarian batas. Metode ini mencapai hasil yang kompetitif dengan menyerang mekanisme pertahanan populer, sementara efisiensi kuerinya perlu ditingkatkan.

Li dkk. menunjukkan bahwa banyaknya jumlah perulangan kueri untuk serangan berbasis batas disebabkan oleh masukan berdimensi tinggi (misalnya gambar). Dengan demikian, tiga metode optimasi subruang (yaitu, subruang spasial, subruang frekuensi, dan subruang komponen utama) dieksplorasi dalam Serangan Blackbox Berbasis Batas Efisien Kueri (QEBA) untuk pengambilan sampel perturbasi. Secara khusus, subruang spasial memanfaatkan interpolasi linier untuk mereduksi gambar menjadi ruang berdimensi rendah. Subruang frekuensi kedua diperoleh dengan transformasi kosinus diskrit (DCT), sedangkan subruang ketiga memilih komponen utama dengan analisis komponen utama (PCA).

Serangan Berbasis Transformasi

Terakhir, serangan berbasis transformasi menghasilkan gambar manipulasi dengan menggeser lokasi spasial piksel alih-alih mengubah nilainya secara langsung. Misalnya, Xiao dkk. mengusulkan metode manipulasi yang ditransformasikan secara spasial (stAdv), yang mengukur besarnya gangguan melalui distorsi geometri lokal dan bukan norma Lp. Alasannya adalah bahwa transformasi spasial pada suatu gambar sering kali menyebabkan hilangnya Lp dalam jumlah besar, namun gangguan tersebut secara visual tidak terlihat oleh manusia dan sulit untuk dipertahankan. Untuk setiap piksel, lokasi spasialnya dapat dipindahkan ke tetangga empat piksel, yaitu kiri atas, kanan atas, kiri bawah, dan kanan bawah. StAdv membangun fungsi tujuan untuk meminimalkan distorsi lokal dan menyelesaikan masalah minimalisasi ini dengan pengoptimal L-BFGS.

Engstrom dkk. juga menemukan bahwa memutar atau menerjemahkan gambar alami saja sudah cukup untuk mengelabui model penglihatan mendalam. Untuk melakukan translasi dan rotasi secara bersamaan, penulis mendefinisikan tiga parameter dimana dua parameter untuk translasi dan satu parameter sudut untuk mengontrol rotasi. Kemudian mereka merancang tiga cara berbeda untuk mengoptimalkan parameter ini, termasuk metode orde pertama, pencarian grid, dan pemilihan terburuk-of-k. Metode orde pertama memerlukan pengetahuan penuh tentang pengklasifikasi untuk menghitung gradien fungsi

kerugian, sedangkan strategi kedua dan ketiga dapat dilakukan dalam pengaturan kotak hitam.

Wang dkk. menyelidiki pengaruh transformasi spasial gambar pada tugas penerjemahan gambar-ke-gambar (Im2Im), yang lebih canggih daripada masalah klasifikasi murni. Mereka mengungkapkan bahwa transformasi gambar geometris (yaitu translasi, rotasi, dan skala) di domain masukan dapat menyebabkan peta warna kerangka Im2Im yang salah di domain target. Berbeda dengan karya-karya sebelumnya yang hanya mengandalkan transformasi spasial, Chen dkk. mengintegrasikan transformasi spasial linier (yaitu, transformasi affine) dengan transformasi warna dan mengusulkan serangan kombinasi dua fase. Kecuali transformasi affine, penulis mendefinisikan transformasi warna sebagai perubahan iluminasi, karena penyesuaian ini tidak mengubah informasi semantik suatu gambar. Selain itu, karena norma Lp tidak sesuai untuk mengukur kualitas manipulasi dalam serangan transformasi, penulis menggunakan indeks kesamaan struktural (SSI) untuk mengukur kualitas persepsi. Model musuh ini berpotensi diterapkan untuk melindungi interaksi pengguna sosial untuk pembelajaran pengaruh.

Berdasarkan pengetahuan penyerang, metode ini dapat dibagi menjadi serangan white-box, serangan blackbox, dan serangan grey-box. Secara khusus, serangan kotak putih mengasumsikan penyerang mengetahui segala sesuatu tentang model korban (misalnya arsitektur, parameter, metode pelatihan, dan data), serangan kotak hitam mengasumsikan musuh hanya mengetahui keluaran model (label prediksi atau probabilitas) jika diberi masukan, dan serangan kotak abu-abu berarti skenario di mana peretas mengetahui sebagian informasi (misalnya struktur jaringan) dan sisanya (misalnya parameter) hilang. Berdasarkan kekhususan penyerang, metode ini termasuk dalam serangan bertarget dimana model mengeluarkan label yang ditentukan pengguna, atau serangan tidak bertarget dimana model disesatkan ke label lain selain label yang benar. Ringkasannya disajikan pada Tabel 6.1.

Tabel 6.1 Ringkasan properti untuk berbagai metode penyerangan. Properti tersebut adalah Serangan bertarget, Serangan tidak bertarget, Serangan kotak putih, Serangan kotak hitam, dan Serangan kotak abu-abu

Metode menyerang	Properti				
	Ditargetkan	Tidak ditargetkan	Putih	Hitam	abu-abu
FGSM	✓	✓	✓		
I-FGSM	✓	✓	✓		
DeepFool		✓	✓		
MI-FGSM	✓	✓	✓	✓	
Xiang et al.	✓				✓
Narodytska et al.	Top-k misclass		✓		
Hayes and Danezis	✓	✓		✓	
Ilyas et al.	✓			✓	
Zhao et al. Boundary Attack	✓	✓ ✓	✓	 ✓	
Pointwise attack	✓	✓		✓	
Boundary Attack ++	✓	✓		✓	

HSJA	✓	✓		✓	
QEBA	✓			✓	
stAdv	✓		✓		
Engstrom et al.		✓		✓	
Wang et al.		✓		✓	
Chen et al.	✓		✓		

6.2 SERANGAN MANIPULASI TERHADAP TEKS

Dibandingkan dengan serangan gambar yang merugikan, kerentanan model pembelajaran mendalam dalam pengenalan teks masih terlalu diremehkan. Ada beberapa kesulitan dalam menyusun sampel teks pembelajaran mesin adversarial. Pertama, keluaran sistem serangan teks harus memenuhi berbagai sifat alami, seperti kebenaran leksikal, kebenaran sintaksis, dan kesamaan semantik. Properti ini memastikan prediksi manusia tidak akan berubah setelah serangan musuh. Kedua, kata-kata dalam rangkaian teks adalah token terpisah, bukan nilai piksel kontinu dalam ruang gambar. Oleh karena itu, tidak mungkin menghitung gradien model secara langsung terhadap setiap kata. Metode bundaran langsung memetakan kalimat-kalimat ke dalam ruang penyematan kata yang berkesinambungan, namun metode ini tidak dapat memastikan bahwa kata-kata yang ditutup dalam ruang penyematan tersebut koheren secara sintaksis bagi pembaca. Ketiga, membuat gangguan kecil pada banyak piksel masih dapat menghasilkan gambar yang bermakna menurut persepsi manusia. Namun, perubahan kecil apa pun, bahkan satu kata pun, pada dokumen teks dapat membuat sebuah kalimat menjadi tidak bermakna.

Upaya pertama serangan teks dapat ditelusuri kembali ke tahun 2016, ketika Papernot dkk. menyelidiki ketahanan jaringan saraf berulang (RNN) dalam memproses data sekuensial. Dalam karya ini, Papernot dkk. membuktikan bahwa RNN dapat ditipu 100% dengan mengubah rata-rata 9 kata dalam ulasan film 71 kata untuk tugas analisis sentimen. Sejak tahun 2016, beberapa baris karya telah diusulkan untuk menghasilkan contoh teks manipulasi , termasuk serangan tingkat karakter, serangan tingkat kata, dan serangan tingkat kalimat serangan Tabel 6.2 menguraikan tiga contoh manipulasi yang dihasilkan oleh strategi serangan yang berbeda. Secara khusus, serangan tingkat karakter menghasilkan teks manipulasi dengan menghapus, menyisipkan, atau menukar karakter. Namun, modifikasi tingkat karakter ini menyebabkan kata-kata salah eja, yang dapat dengan mudah dideteksi oleh mesin pemeriksa ejaan.

Serangan tingkat kalimat menggabungkan kalimat manipulasi sebelum atau sesudah teks asli untuk membingungkan model arsitektur yang mendalam, namun biasanya menyebabkan perubahan semantik yang dramatis dan menghasilkan kalimat yang tidak dapat dipahami manusia. Untuk mengatasi kelemahan ini, sebagian besar penelitian terbaru berfokus pada serangan tingkat kata, yang menggantikan kata asli dengan kata lain yang dipilih dengan cermat. Namun, strategi substitusi kata yang ada masih jauh dari sempurna dalam mencapai tingkat keberhasilan serangan yang tinggi dan tingkat substitusi yang rendah. Pada bagian ini, kami meninjau metode serangan teks terkait, termasuk serangan tingkat karakter, serangan tingkat kalimat, serangan tingkat kata, dan serangan bertingkat.

Tabel 6.2 Tiga contoh teks adversarial yang berhasil dihasilkan oleh strategi serangan tingkat karakter, serangan tingkat kalimat, dan serangan tingkat kata

Serangan tingkat karakter mengubah karakter input dari "p → B" .
Asli: Kanselir Gordon Brown berusaha meredam spekulasi mengenai siapa yang seharusnya memimpin Partai Buruh dan mengalihkan serangan terhadap oposisi Konservatif.
Manipulasi : Kanselir Gordon Brown berusaha meredam spekulasi mengenai siapa yang harus memimpin Partai Buruh dan mengarahkan serangan terhadap Partai Konservatif oBposition.
Serangan tingkat kalimat menambahkan satu kalimat di akhir masukan
Asli: Peyton Manning menjadi gelandang pertama yang memimpin dua tim berbeda ke beberapa Super Bowl. Dia juga gelandang tertua yang pernah bermain di Super Bowl pada usia 39. Rekor sebelumnya dipegang oleh John Elway, yang memimpin Broncos meraih kemenangan di Super Bowl XXXIII pada usia 38 dan saat ini menjabat sebagai Wakil Presiden Eksekutif Operasi Sepak Bola dan Denver. Manajer umum
Musuh: Peyton Manning menjadi quarterback pertama yang memimpin dua tim berbeda ke beberapa Super Bowl. Dia juga gelandang tertua yang pernah bermain di Super Bowl pada usia 39. Rekor sebelumnya dipegang oleh John Elway, yang memimpin Broncos meraih kemenangan di Super Bowl XXXIII pada usia 38 dan saat ini menjabat sebagai Wakil Presiden Eksekutif Operasi Sepak Bola dan Denver. Manajer umum. Quarterback Jeff Dean memiliki nomor punggung 37 di Champ Bowl XXXIV
Serangan tingkat kata menggantikan kata masukan dari "f unny → menggelikan"
Asli: Ah, film ini lucu sekali, namun aneh. Saya suka bagaimana mereka mempertahankan bahasa Shakespeare dalam film ini, rasanya ironis karena betapa bodohnya film tersebut. film ini pasti menjadi salah satu film terbaik troma. sangat direkomendasikan untuk kesenangan yang tidak masuk akal!
Adversarial: Ya ampun, film ini menggelikan sekali, namun aneh. Saya suka bagaimana mereka mempertahankan bahasa Shakespeare dalam film ini, rasanya ironis karena betapa bodohnya film tersebut. film ini pasti menjadi salah satu film terbaik troma. sangat direkomendasikan untuk kesenangan yang tidak masuk akal!

Serangan Tingkat Karakter

Pertama, serangan tingkat karakter menghasilkan teks manipulasi dengan menghapus, menyisipkan, atau menukar karakter. Belinkov dan Bisk merancang empat jenis derau sintetik: (1) menukar dua karakter yang berdekatan tetapi mengecualikan huruf pertama dan terakhir (misalnya, noise nosie), (2) mengacak urutan semua huruf dalam sebuah kata kecuali untuk pertama dan terakhir (misalnya, noise nisoe), (3) kata yang sepenuhnya acak termasuk karakter pertama dan terakhir (misalnya, noise iones), dan (4) kesalahan ketik keyboard yang secara acak mengganti satu huruf di setiap kata dengan tombol yang berdekatan (misalnya, , kebisingan tidak ada). Strategi ini sebagian besar dapat menyesatkan model terjemahan mesin saraf (NMT). Namun, mereka memodifikasi setiap kata dalam kalimat masukan semampu mereka, sehingga menyebabkan hilangnya gangguan yang tinggi. Misalnya, "tukar" dua huruf diterapkan pada semua kata dengan panjang 4, karena tidak mengubah huruf pertama dan terakhir.

Untuk mengurangi derajat distorsi, Ebrahimi et al. mengusulkan HotFlip, yang merepresentasikan setiap karakter sebagai vektor one-hot dan mengusulkan dua operasi

karakter, yaitu penyisipan karakter dan penghapusan karakter. Secara khusus, HotFlip memperkirakan perubahan karakter terbaik (alias, operasi flip atom) dengan menghitung turunan arah sehubungan dengan representasi vektor one-hot. Kemudian ia menggunakan pencarian sinar untuk menemukan serangkaian manipulasi yang dapat bekerja dengan baik untuk membingungkan pengklasifikasi yang terlatih. Selain itu, HotFlip menetapkan batas atas pembalikan karakter sebesar 20% untuk setiap sampel pelatihan guna membatasi manipulasi.

Untuk meminimalkan jarak edit dan mengurangi tingkat distorsi, Gao et al. merancang DeepWordBug kotak hitam dan membuat gangguan teks hanya pada kata-kata yang paling penting. Secara khusus, ini mengevaluasi skor kepentingan kata dengan langsung menghapus kata satu per satu dan membandingkan perubahan prediksi. DeepWordBug memodifikasi kata dengan mengikuti empat operasi karakter, termasuk (1) mengganti huruf dalam kata dengan huruf acak, (2) menghapus karakter acak dalam kata, (3) memasukkan huruf acak ke dalam kata, dan (4) menukar dua huruf yang berdekatan dalam sebuah kata. Mereka mendefinisikan jarak edit sebagai jarak Levenshtein, sehingga jarak edit untuk (1), (2), dan (3) adalah 1, tetapi jarak untuk (4) adalah 2.

Gil dkk. menunjukkan bahwa metode HotFlip yang dirancang dengan pengaturan kotak putih dapat diterapkan untuk melakukan serangan kotak hitam melalui distilasi yang efisien. Prosedur putih-ke-hitam ini berisi tiga langkah: pertama, melatih model klasifikasi teks sumber dan model kotak hitam target; kedua, buat contoh manipulasi dengan menyerang model sumber dengan HotFlip di bawah kotak putih; dan ketiga, melatih penyerang untuk menghasilkan contoh manipulasi baru untuk menyerang model target kotak hitam. Penyerang dilatih menggunakan pasangan (input, output) dengan fungsi kerugian cross-entropy yang dirancang dengan cermat, di mana input menunjukkan kata input asli dan outputnya adalah modifikasi yang dilakukan pada langkah kedua.

Eger dkk. mengusulkan algoritma Visual Perturber (VIPER) untuk mengganti karakter dengan simbol yang mirip secara visual, yang biasa digunakan dalam bahasa gaul Internet (misalnya, n00b) dan komentar beracun (misalnya, !d10t), dll. Keuntungan dari serangan visual meliputi tidak diperlukan pengetahuan linguistik apa pun yang melampaui tingkat karakter dan tidak terlalu merusak persepsi dan pemahaman manusia. Kandidat simbol yang mirip secara visual dipilih dari tiga ruang penyematan karakter, termasuk ruang penyematan karakter berbasis gambar (ICES), ruang penyematan karakter berbasis deskripsi (DCES), dan ruang penyematan karakter mudah (ECES). ECES mencapai efek maksimal pada model target dengan menambahkan simbol di bawah atau di atas karakter (misalnya, c c), namun gangguan ini memerlukan pemilihan manual. Namun, satu kelemahan umum serangan tingkat karakter adalah bahwa serangan tersebut merusak batasan leksikal dan menyebabkan kata salah eja, yang dapat dengan mudah dideteksi dan dihapus oleh mesin pemeriksa ejaan yang dipasang sebelum pengklasifikasi.

Serangan Tingkat Kalimat

Serangan tingkat kalimat menggabungkan kalimat manipulasi sebelum atau lebih umum setelah teks masukan bersih untuk membingungkan model arsitektur mendalam.

Misalnya, Jia dan Liang menambahkan kalimat yang kompatibel ke akhir paragraf untuk mengelabui model pemahaman bacaan (RCM). Kalimat manipulasi terlihat mirip dengan pertanyaan awal dengan menggabungkan pertanyaan yang diubah dan jawaban palsu, yang bertujuan untuk menyesatkan RCM ke lokasi jawaban yang salah. Namun demikian, strategi ini memerlukan banyak campur tangan manusia dan tidak dapat sepenuhnya diotomatisasi, misalnya, strategi ini bergantung pada sekitar 50 aturan yang ditentukan secara manual untuk memastikan kalimat manipulasi dalam bentuk deklaratif.

Wallace dkk. mencari pemicu manipulasi universal, yaitu urutan input-agnostik, yang menyebabkan prediksi target spesifik ketika digabungkan ke input apa pun dari kumpulan data yang sama. Urutan universal diinisialisasi secara acak dan diperbarui secara berulang untuk meningkatkan kemungkinan prediksi target menggunakan gradien penggantian token sebagai HotFlip, sementara metode ini gagal menjamin keluaran yang bermakna secara semantik bagi persepsi manusia dan sering kali menghasilkan teks tidak beraturan (misalnya, “zonasi penyadapan fiennes”).

Baru-baru ini, Song dkk. mengusulkan Natural Universal Trigger Search (NUTS) untuk membuat pemicu lancar yang membawa makna semantik. NUTS menggunakan autoencoder adversarially regularized (ARAE) yang telah dilatih sebelumnya untuk menghasilkan pemicu dan mengadopsi pencarian berbasis gradien untuk memaksimalkan fungsi kerugian pada sistem klasifikasi. Selama optimasi, beberapa vektor kebisingan independen (256 vektor dalam percobaannya) diinisialisasi terlebih dahulu. Kemudian kandidat pemicu yang dioptimalkan tersebut diberi peringkat ulang berdasarkan akurasi pengklasifikasi dan kealamian.

Wang dkk. mengusulkan model Pembuatan Teks pembelajaran mesin Adversarial Terkendali (CATGen) yang menghasilkan kalimat manipulasi dengan mengubah atribut kalimat masukan yang dapat dikontrol. Untuk lebih spesifiknya, CATGen berisi dua modul utama, yaitu kerangka encoder-decoder untuk pembuatan teks dan pengklasifikasi atribut. Encoder dan decoder keduanya merupakan RNN untuk mempelajari salinan kalimat masukan. Pengklasifikasi atribut dilatih pada kumpulan data tambahan, yang bertujuan untuk mempelajari atribut yang dapat dikontrol (misalnya kategori, jenis kelamin, domain) yang tidak relevan dengan label tugas (misalnya positif, negatif). Misalnya dengan mengubah atribut dari “Kitchen” menjadi “Phone”, kalimat input “pisau yang luar biasa, sudah lama digunakan untuk edc saya, hanya diganti karena bosan dengan pisau yang itu-itu saja (Pos.)” menjadi “ kasus yang luar biasa. sudah lama digunakan di iPhone5 saya, hanya masalah karena saya bosan dengan Kindle lama yang itu-itu saja (Neg.).”

Kecuali untuk tugas klasifikasi, Han et al. menyelidiki serangan manipulasi untuk tugas prediksi terstruktur di NLP dan mengusulkan generator kalimat sequence-to-sequence (seq2seq). Berbeda dengan tugas klasifikasi, salah satu tantangan khusus untuk model prediksi terstruktur adalah keluaran terstruktur dari model prediksi lebih sensitif terhadap gangguan kecil pada kalimat masukan. Misalnya, menggeser hanya satu kata dari kalimat “*Saya seorang penulis*” menjadi “*Saya memecat seorang penulis*” akan membuat pengurai ketergantungan menghasilkan pohon penguraian yang berbeda. Untuk mengatasi masalah

ini, generator seq2seq dilatih dengan pembelajaran penguatan dan mengambil pohon parse sebagai salah satu istilah dari fungsi penghargaan. Kemudian dapat digunakan untuk menghasilkan kalimat manipulasi secara langsung dengan memberikan kalimat masukan tanpa mengakses model korban, yaitu bertindak sebagai penyerang online.

Le dkk. mengeksplorasi kekokohan model deteksi berita palsu saraf dan mengusulkan kerangka kerja Malcom untuk menghasilkan komentar manipulasi . Berita palsu biasanya terdiri dari judul, isi, komentar, dan balasan, dimana judul dan isi tersebut tidak dapat diubah jika penyerangnya bukan penerbitnya. Namun, musuh dapat memberikan komentar jahat apa pun untuk artikel yang diterbitkan. Sebagai bagian dari masukan, komentar musuh dapat menyesatkan detektor yang sama tanpa kepemilikan artikel target. Untuk memastikan komentar musuh relevan dengan artikel, Malcom melatih generator teks bersyarat bersama dengan modul STYLE dan modul ATTACK dengan merancang fungsi tujuan di bawah kotak putih. Serangan kotak putih ini juga dapat ditransfer ke beberapa pengklasifikasi berita palsu yang tidak diketahui di bawah pengaturan kotak hitam.

Selain itu, serangan tingkat kalimat biasanya muncul di tugas NLP lainnya, seperti terjemahan mesin alami (NMT) dan menjawab pertanyaan (QA). Namun, karena metode ini memanipulasi dokumen teks pada tingkat kalimat, metode ini biasanya menimbulkan biaya gangguan yang tinggi dan perubahan semantik yang signifikan.

Serangan Tingkat Kata

Serangan tingkat kata menggantikan kata masukan asli dengan kata yang dipilih dengan cermat. Permasalahan pokoknya adalah (1) bagaimana memilih calon kata yang tepat dan (2) bagaimana menentukan urutan substitusi kata. Baru-baru ini, Papernot dkk. memproyeksikan kata-kata ke dalam ruang penyematan 128 dimensi dan memanfaatkan matriks Jacobian untuk mengevaluasi interaksi input-output. Namun, gangguan kecil pada ruang penyematan dapat menyebabkan kata-kata yang sama sekali tidak relevan karena tidak ada jaminan pasti bahwa kata-kata yang dekat dengan ruang penyematan memiliki kesamaan semantik. Oleh karena itu, penelitian selanjutnya berfokus pada strategi substitusi sinonim yang mencari sinonim dari ruang penyematan GloVe, tesaurus yang ada (misalnya, WordNet dan HowNet), atau model bahasa bertopeng (MLM) BERT.

Dengan menggunakan GloVe, Alzantot dkk. merancang algoritma genetika berbasis populasi (GA) untuk meniru seleksi alam. Prosedur optimasi dimulai dari generasi awal dengan serangkaian modifikasi kata yang berbeda. Di setiap generasi berikutnya, persilangan dan mutasi digunakan untuk evolusi populasi dan optimalisasi kandidat. Khususnya, persilangan memerlukan lebih dari satu solusi induk untuk menghasilkan satu solusi anak, dan mutasi dirancang untuk meningkatkan keragaman anggota populasi. Jin dkk. menyajikan TextFooler, yang mengumpulkan kandidat pengganti dari ruang penyematan GloVe. Berbeda dengan GA, TextFooler menentukan urutan substitusi kata dengan menghitung skor kepentingan kata (WIS). Secara khusus, WIS didefinisikan sebagai pengurangan probabilitas label yang benar dan peningkatan skor label yang salah dengan menghapus setiap kata masukan secara berulang. Namun, penyematan GloVe biasanya gagal membedakan antonim dari sinonim. Misalnya, tetangga terdekat untuk mahal di ruang GloVe adalah pricey, cheap,

costly, di mana cheap adalah antonimnya. Oleh karena itu, algoritma berbasis Glove harus menggunakan metode counter-fitting pada vektor musuh postprocess untuk memastikan batasan semantik.

Dibandingkan dengan GloVe, menggunakan tesaurus linguistik yang terorganisir dengan baik, misalnya WordNet berbasis sinonim dan HowNet berbasis sememe, merupakan cara yang lebih mudah. Secara khusus, WordNet adalah kumpulan data leksikal bahasa Inggris yang besar, di mana kata benda, kata kerja, kata sifat, dan kata keterangan dikelompokkan ke dalam kumpulan sinonim kognitif (synsets). HowNet memberi anotasi pada kata-kata berdasarkan sememnya, di mana sememe adalah unit minimum makna semantik dalam linguistik. Ren dkk. mencari sinonim untuk setiap kata masukan dari synset WordNet dan menentukan prioritas penggantian kata masukan dengan menghitung probabilitas arti-penting kata tertimbang (PWWS). Kemudian mereka secara berurutan mengganti setiap kata dengan kandidat terbaik mengikuti urutan PWWS hingga menemukan sampel pembelajaran mesin adversarial yang berhasil. Zang dkk. menyatakan bahwa HowNet berbasis sememe dapat menyediakan lebih banyak kata pengganti daripada WordNet dan mengusulkan optimasi gerombolan partikel (PSO) untuk menentukan kelompok kata mana yang harus diserang. Dalam PSO, setiap kalimat diperlakukan sebagai partikel dalam ruang pencarian, dan setiap dimensi partikel berhubungan dengan sebuah kata. Oleh karena itu, contoh manipulasi yang berhasil dapat ditemukan dengan mengoptimalkan lokasi partikel secara bertahap.

Beberapa penelitian terbaru menggunakan model bahasa bertopeng BERT (MLM) untuk menghasilkan gangguan kontekstual, seperti BERT-Attack dan contoh pembelajaran mesin adversarial berbasis BERT (BAE). BERT MLM yang telah dilatih sebelumnya dapat memastikan token yang diprediksi sesuai dengan kalimat dengan baik tetapi tidak dapat mempertahankan kesamaan semantik. Misalnya, dalam kalimat “makanannya adalah [MASK]”, memprediksi [MASK] sebagai baik atau buruk adalah hal yang sama namun menghasilkan label sentimen yang berlawanan. Selain itu, BERT-Attack dan BAE mengadopsi urutan penggantian kata statis yang dipandu oleh skor kepentingan kata (WIS), yang menyebabkan substitusi kata redundansi. Perbedaannya terletak pada Garg dan Ramakrishnan yang mendefinisikan WIS sebagai penurunan probabilitas label yang benar setelah menghapus sebuah kata, sedangkan Li et al. mengganti setiap kata aslinya dengan simbol tiruan [MASK].

Selain itu, Li dkk. mempresentasikan model Contoh AdversaRial Kontekstual (CLARE) untuk menghasilkan keluaran manipulasi yang lancar melalui prosedur mask-then-infill. Alih-alih menggunakan BERT MLM, CLARE menggunakan RoBERTa MLM yang telah dilatih sebelumnya untuk memberikan kata-kata pengisi yang dikontekstualisasikan. CLARE mengadopsi tiga gangguan teks, yaitu penggantian, penyisipan, dan penggabungan, yaitu mengganti token masukan, memasukkan token baru, dan menggabungkan bigram. Untuk setiap kata masukan, CLARE akan mencoba ketiga gangguan ini dan memilih salah satu yang meminimalkan kemungkinan label emas.

Serangan Bertingkat

Serangan bertingkat menggabungkan setidaknya dua dari tiga strategi serangan di atas untuk membuat teks manipulasi . Berbeda dengan strategi tunggal, algoritma serangan bertingkat relatif lebih rumit dan mahal secara komputasi. Misalnya, Liang dkk. disajikan untuk mendandani masukan teks pada tingkat karakter dan tingkat kata melalui tiga strategi, yaitu penyisipan, modifikasi, dan penghapusan. Strategi ini diterapkan pada karakter-karakter populer dan kata-kata populer (yaitu, item-item penting klasifikasi) yang diidentifikasi dengan memanfaatkan gradien biaya. Selain itu, mereka mengusulkan teknik watermarking bahasa alami untuk meningkatkan keterbacaan dan kegunaan teks manipulasi , misalnya menyisipkan frasa kosong secara semantik. Perlu disebutkan bahwa menggunakan satu strategi (misalnya, penghapusan) sering kali tidak cukup untuk mengelabui pengklasifikasi dan menggabungkan tiga strategi sangat penting untuk menyusun sampel manipulasi yang halus. Namun, prinsip optimasi yang jelas tentang cara menggabungkan strategi ini masih kurang.

Li dkk. mengusulkan TextBugger yang memodifikasi teks jinak pada tingkat kata dan tingkat karakter. Secara khusus, ini mendefinisikan lima jenis metode gangguan bug, termasuk (1) menyisipkan spasi ke dalam kata, (2) menghapus karakter acak dari kata tersebut kecuali karakter pertama dan terakhir, (3) menukar dua huruf yang berdekatan dari sebuah kata, (4) mengganti karakter dengan karakter yang serupa secara visual, dan (5) mengganti sebuah kata dengan k tetangga terdekatnya di ruang penyematan GloVe. Untuk setiap kata masukan, ia memilih bug terbaik dari lima strategi ini sebagai salah satu yang paling mengurangi kemungkinan kebenaran dasar. Keluaran akhir manipulasi dibuat dengan mengulangi prosedur ini secara berulang pada setiap kata masukan.

Wang dkk. menyajikan kerangka serangan berbasis pohon T3 yang mengganggu teks pada tingkat kata (T3(WORD)) dan tingkat kalimat (T3(SENT)). Komponen inti T3 adalah autoencoder berbasis pohon terlatih, yang dapat mengubah ruang teks diskrit menjadi ruang penyematan semantik berkelanjutan. Hal ini memecahkan tantangan masukan diskrit sehingga metode pengoptimalan berbasis gradien dapat digunakan untuk menemukan penyematan yang berlawanan. Terakhir, penyematan adversarial dapat dipetakan kembali ke teks adversarial dengan decoder berbasis pohon dengan seperangkat aturan tata bahasa pohon. Tingkat keberhasilan serangan yang tinggi dicapai melalui serangkaian iterasi. Mirip dengan serangan gambar, ciri khas metode serangan teks ini dirangkum dalam Tabel 6.3. Seperti yang bisa kita lihat dari Tabel 6.3 bahwa sebagian besar metode serangan teks yang ada dirancang untuk serangan tidak bertarget.

Tabel 6.3 Ringkasan properti untuk berbagai metode serangan teks. Propertinya adalah Serangan bertarget, Serangan tidak bertarget, Serangan kotak putih, dan Serangan kotak hitam

Metode menyerang	Properti			
	Ditargetkan	Tidak ditargetkan	Putih	Hitam
Belinkov and Bisk		✓		✓
HotFlip		✓	✓	

DeepWordBug		✓		✓
Gil et al		✓	✓	✓
VIPER		✓		✓
Jia and Liang		✓		✓
Wallace et a	✓		✓	
NUTS		✓	✓	
CATGen et al.		✓	✓	
Han et al		✓		✓
Malcom	✓		✓	✓
Papernot et al		✓	✓	
Alzantot et al	✓			✓
TextFooler		✓		✓
PWWS		✓		✓
PSO	✓			✓
BERT-Attack		✓		✓
BAE		✓		✓
CLARE		✓		✓
Liang et al	✓		✓	✓
TextBugger		✓	✓	✓
T3	✓		✓	

6.3 PENYARINGAN SPAM

Spam Teks

Pemfilteran spam email telah dianalisis sebagai masalah pembelajaran malas dalam penyimpangan konsep. Kazemian dkk. membandingkan teknik pembelajaran mesin untuk mendeteksi halaman web berbahaya. Contoh manipulasi dalam sistem pemahaman bacaan dianalisis oleh Jia et al. Chen dkk. membahas contoh manipulasi dalam pengklasifikasi pembelajaran mesin untuk deteksi malware. Miyato dkk. membahas pelatihan manipulasi dengan contoh manipulasi pada penyematan kata dalam jaringan saraf berulang. Dasgupta dkk. skenario serangan manipulasi dalam klasifikasi teks untuk analisis sentimen di situs media sosial. Cheng dkk. membuat contoh manipulasi untuk model urutan-ke-urutan (seq2seq). Metode kami tidak spesifik untuk sumber data tertentu. Kami bereksperimen dengan database gambar, database teks, dan database deret waktu.

Spam Gambar

Masalah deteksi spam gambar adalah bagian dari pemfilteran data multimedia berbasis konten di lingkungan yang bermusuhan. Data multimedia semacam itu sering kali dihasilkan di komunitas Internet dan jaringan seluler. Menurut survei Attar dkk., spam gambar dibuat dengan menyematkan pesan teks spam ke dalam gambar. Tujuan manipulasi nya adalah untuk mencegah pengenalan teks oleh perangkat lunak pengenalan karakter optik. Deteksi kata kunci, kategorisasi teks, klasifikasi gambar, dan deteksi hampir duplikat adalah teknik yang ada untuk mendeteksi spam gambar. Dalam menerapkan teknik ini pada pembelajaran mesin adversarial, asumsi yang mendasarinya adalah bahwa teks dalam gambar yang sah (dan fitur terkait yang membedakan antara gambar spam dan gambar yang sah) tidak mungkin dikaburkan dengan fitur adversarial.

Fitur manipulasi juga dapat dibangun berdasarkan alasan pengambilan gambar berbasis konten di mana pencarian gambar spam atau gambar sah didorong oleh serangkaian fitur tingkat rendah yang ditemukan dalam gambar kueri. Dalam metode seperti itu, jarak antara gambar kueri dan templat dalam database dihitung untuk setiap ruang fitur dan dibandingkan dengan ambang batas untuk memutuskan apakah suatu gambar merupakan gambar spam atau gambar sah. Oleh karena itu, kemampuan generalisasi algoritma pembelajaran mesin adversarial terhadap spam gambar sangat bergantung pada pilihan fitur yang tepat untuk manipulasi data adversarial. Dalam literatur yang ada, pilihan fitur bergantung pada asumsi tentang properti yang paling membedakan antara gambar spam dan gambar sah. Fitur yang paling umum digunakan mencakup kebingungan teks, area teks, properti gambar tingkat rendah (seperti warna, tekstur, dll.), kesamaan gambar, kesamaan wilayah gambar, dan metadata gambar. Fitur-fitur yang relevan kemudian dipilih berdasarkan hasil mengenai akurasi klasifikasi, rasio positif sebenarnya, rasio positif palsu, presisi, dan perolehan kembali. Pengklasifikasi yang paling umum mencakup mesin vektor dukungan, pohon keputusan, model entropi maksimum, dan jaringan Bayesian.

Spam Biometrik

Biometrik adalah bidang penelitian yang mengutamakan keamanan. Keamanan dalam biometrik ditentukan oleh kerentanan metode klasifikasi pola. Biggio dkk. menyelidiki serangan dan pertahanan untuk pembelajaran manipulasi dalam sistem biometrik adaptif. Serangan dalam sistem biometrik adaptif berkaitan dengan pengenalan biometrik atau perubahan sifat biometrik seiring waktu. Untuk menangani serangan ini secara efektif, templat biometrik yang disimpan harus sesuai dengan identitas yang diklaim yang dikirimkan selama verifikasi.

Titik-titik serangan yang ditemukan selama proses pencocokan identitas biometrik dikategorikan oleh Biggio et al. seperti input sensor, ekstraksi fitur, database template, algoritma pencocokan, pembaruan template, aturan penilaian, dan ambang batas penilaian. Dalam biometrik adaptif, titik serangan tambahan mencakup pencurian template dan infeksi malware tanpa memperhatikan kegagalan intrinsik. Serangan-serangan ini selanjutnya diklasifikasikan menjadi serangan terhadap sensor, antarmuka dan saluran yang menghubungkan modul, modul pemrosesan dan algoritme, serta database templat.

Serangan berikut terlihat dalam sistem biometrik adaptif:

- Serangan spoofing membuat ciri biometrik palsu untuk menyamar sebagai klien terdaftar.
- Serangan ulang menampilkan biometrik yang dicuri sebagai fitur dalam algoritma pencocokan.
- Serangan pendakian bukit mempengaruhi saluran komunikasi dengan mengirimkan data yang terganggu secara berulang ke algoritma pencocokan dan menyimpan data yang memberikan skor pencocokan maksimum. Iterasi dalam serangan berlanjut hingga konvergensi metode optimasi yang digunakan oleh musuh.
- Serangan infeksi malware mengeksploitasi kerentanan perangkat lunak dan perangkat keras yang diketahui melalui teknik peretasan dan praktik pemrograman.

- Serangan pencurian template menargetkan database template yang tidak dilindungi dengan benar dan tidak dienkripsi.

Biggio dkk. kemudian melanjutkan untuk mengkarakterisasi serangan dalam sistem biometrik sesuai dengan kerangka keamanan yang dibahas oleh Vidyadhari dkk. Skenario serangan spoofing, skenario serangan keracunan, dan skenario serangan penghindaran dibahas sebagai motivasi untuk algoritma pembelajaran manipulasi dalam sistem biometrik yang dirancang dengan aman. Algoritme pencocokan pola dalam sistem biometrik yang aman sesuai desain direkomendasikan untuk dirancang berdasarkan pertimbangan dalam database statistik. Pertimbangan tersebut antara lain pembelajaran dengan invarian, toleransi kesalahan dalam pembelajaran PAC, dan pembelajaran online dengan teori permainan.

BAB 7
GANGGUAN MANIPULATIF UNTUK PERLINDUNGAN PRIVASI

Meskipun contoh konflik (AE) atau gangguan konflik (AP) biasanya diperlakukan sebagai risiko keamanan terkini, contoh tersebut juga dapat berfungsi sebagai alat perlindungan privasi ketika menghadapi serangan privasi berbasis pembelajaran mendalam. Bab ini pertama-tama akan memperkenalkan model privasi untuk data visual, salah satu jenis data terpenting dalam aplikasi pembelajaran mendalam. Kemudian kita akan membahas mekanisme perlindungan privasi berbasis AP yang menggabungkan berbagai tingkat privasi. Meskipun penelitian mengenai topik ini masih dalam tahap awal, bab ini akan mengulas karya-karya mutakhir dan menjelaskan penelitian di masa depan.

7.1 GANGGUAN YANG BERLAWANAN DEMI PELESTARIAN PRIVASI

Karena keakuratannya yang belum pernah terjadi sebelumnya, metode pembelajaran mendalam telah menjadi dasar layanan baru berbasis AI di Internet di era data besar. Sementara itu, hal ini menimbulkan masalah privasi yang jelas. Serangan privasi yang dibantu pembelajaran mendalam dapat mengekstraksi informasi pribadi yang sensitif tidak hanya dari teks tetapi juga dari data tidak terstruktur seperti gambar dan video. Hal ini mendorong kita untuk meninjau kembali tantangan privasi di era big data dengan munculnya berbagai teknologi cerdas. Secara khusus, teknik pembelajaran mendalam yang muncul dapat "secara otomatis mengumpulkan dan memproses jutaan foto atau video untuk mengekstrak informasi pribadi/sensitif dari jejaring sosial." Oleh karena itu, penyelidikan menyeluruh terhadap masalah privasi dalam konteks pembelajaran mendalam merupakan kebutuhan yang mendesak.

Meskipun sebagian besar penelitian yang ada menganggap contoh konflik (AE) atau gangguan konflik (AP) sebagai metode serangan yang mengancam keamanan sistem, AP juga dapat berfungsi sebagai alat perlindungan privasi ketika menghadapi serangan privasi berbasis pembelajaran mendalam. Ide mendasar dari AP adalah untuk menghasilkan gangguan kecil namun disengaja dalam kasus terburuk pada gambar asli, yang menyesatkan model pengenalan berbasis CNN tanpa menyebabkan perbedaan signifikan yang terlihat oleh mata manusia. Oleh karena itu, layak untuk merancang mekanisme perlindungan privasi berbasis AP terhadap serangan privasi.

Ada beberapa penelitian terbaru yang menggunakan gangguan manipulasi sebagai metode perlindungan privasi gambar. Liu dkk. mengusulkan algoritma yang menentang deteksi otomatis menggunakan contoh manipulasi berdasarkan "kerangka kerja RCNN yang Lebih Cepat." Oh dkk. menyiapkan kerangka teori permainan dan mempelajari efektivitas gangguan gambar yang merugikan untuk perlindungan privasi. Shafahi dkk. menyajikan metode berbasis optimasi untuk membuat gambar racun, di mana hanya satu gambar racun yang dapat mengontrol perilaku pengklasifikasi. Jia dkk. mengusulkan kerangka kerja dua fase yang disebut AttriGuard untuk bertahan melawan serangan inferensi atribut yang diluncurkan

oleh pengklasifikasi. Liu dkk. menyelidiki skema penggunaan contoh konflik dalam sistem pembelajaran mesin sehingga skema tersebut tidak dapat mengidentifikasi informasi sensitif dari gambar. Li dkk. mengusulkan untuk menggunakan gangguan manipulasi untuk de-identifikasi wajah.

Komkov dan Petiushko menunjukkan bahwa stiker manipulasi yang diperhitungkan dengan cermat pada topi dapat mengurangi kemungkinan pemakainya dikenali. Zhu dkk. memperkenalkan "serangan polytope" baru di mana gambar racun dirancang untuk mengelilingi gambar target di ruang fitur. Xue dkk. mengusulkan penggunaan gangguan manipulasi untuk melindungi beberapa objek pribadi dalam gambar tampilan jalan. Friedrich dkk. mengusulkan representasi teks medis yang dapat dibagikan dan menjaga privasi untuk pengklasifikasi de-identifikasi. Fawkes membantu pengguna mengenakan "jubah" yang tidak terlihat pada foto mereka sebelum merilisnya. Saat digunakan untuk melatih model pengenalan wajah, gambar "terselubung" ini menghasilkan model fungsional yang secara konsisten menyebabkan gambar normal pengguna salah diidentifikasi.

Karena hampir semua penelitian perlindungan privasi berbasis AP berfokus pada data visual, khususnya data gambar, maka pembahasan dalam bab ini juga akan dilakukan dalam konteks gambar dan video. Pertama-tama kami akan mendefinisikan secara singkat model privasi dalam data visual dan kemudian memperkenalkan tiga kelompok mekanisme perlindungan privasi berbasis AP yang berbeda.

Model Privasi Data Visual

Sebelum kita mulai membahas metode perlindungan privasi, penting untuk terlebih dahulu memperjelas dan memodelkan privasi gambar dan video. Sebagaimana didefinisikan dalam GDPR, privasi didefinisikan sebagai sesuatu yang berkaitan dengan identitas pribadi. Dalam hal ini, model privasi satu tingkat tidak selalu diperlukan, juga tidak cukup untuk sebuah gambar atau video. Misalnya, gambar tampilan jalan yang berisi wajah seseorang bersifat pribadi secara keseluruhan, namun juga berisi banyak informasi non-pribadi. Dalam hal ini, penggunaan privasi tingkat gambar mungkin terlalu kuat untuk penggunaan praktis. Jika kami dapat memastikan wajah tersebut anonim, seluruh gambar dapat digunakan sebagai bagian dari layanan tampilan jalan. Oleh karena itu, lebih umum menggunakan model privasi visual bertingkat. Idenya adalah untuk mendefinisikan model privasi tiga tingkat sebagai berikut:

- Privasi tingkat file: gambar atau video.
- Privasi tingkat objek: wajah, orang, mobil, dll.
- Privasi tingkat fitur: identitas, penampilan, pose, dll.

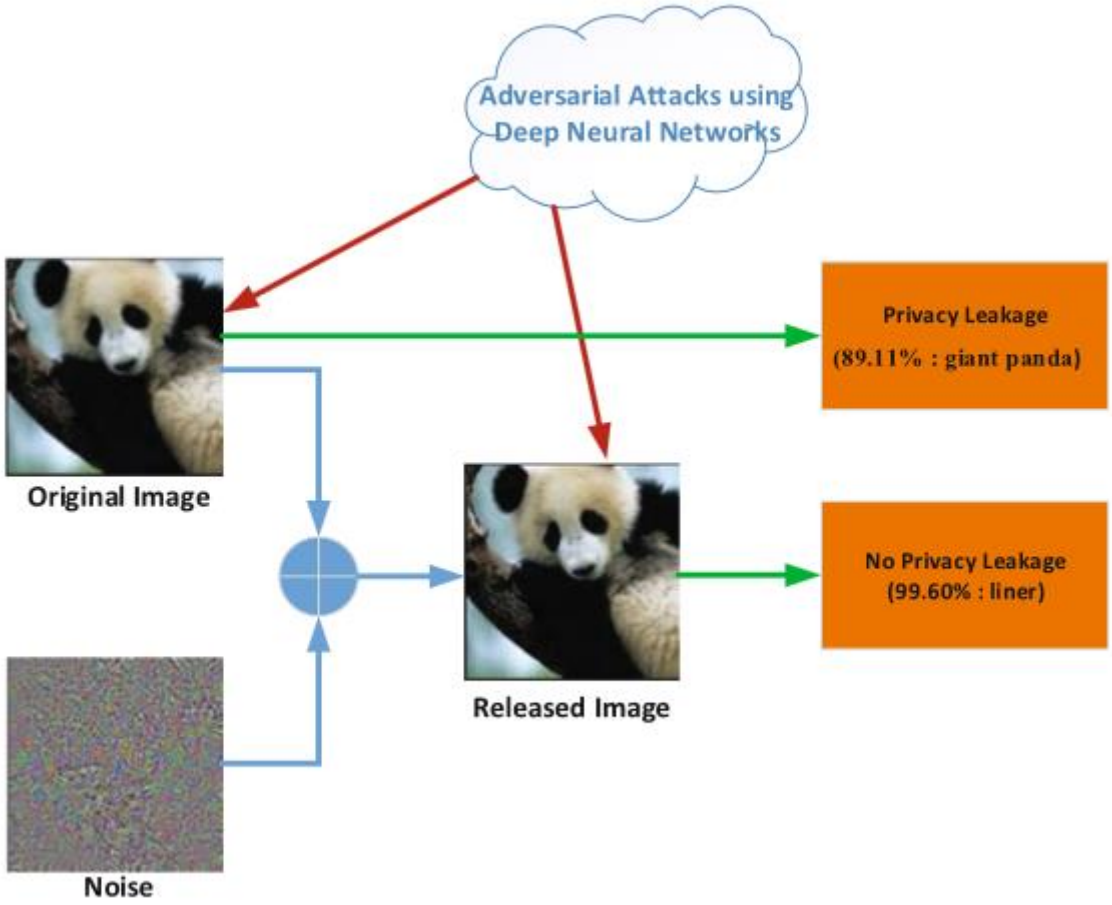
Dari tingkat pertama hingga ketiga, model berubah dari model berbutir kasar menjadi model berbutir halus. Berdasarkan model privasi bertingkat ini, kita dapat membagi berbasis AP yang ada menjadi tiga kelompok dan membahasnya masing-masing di subbagian berikutnya.

7.2 MEKANISME PERLINDUNGAN PRIVASI MELALUI INTERFERENSI LAWAN

Perlindungan Privasi Tingkat File

Untuk perlindungan privasi tingkat file, kami bertujuan untuk menyesatkan alat pembelajaran mendalam ke kelas gambar yang salah. Kami mempertimbangkan skenario jejaring sosial. Lebih detailnya, pengguna memposting gambar di platform jejaring sosial. Misalkan seorang penyerang mengumpulkan gambar melalui crawler dan menggunakan DNN untuk menambang informasi sensitif. Gambar 7.1 menunjukkan contoh arsitektur sistem tersebut. Ketika pengguna membagikan gambar di jejaring sosial tanpa pra-pemrosesan apa pun pada gambar aslinya, musuh yang dilengkapi DNN dapat secara otomatis memperoleh informasi berguna dari gambar ini (yaitu, ini adalah panda raksasa dengan keyakinan tinggi, yang menunjukkan kemungkinan kejadian mengunjungi kebun binatang)'

Informasi sensitif lainnya seperti aktivitas pengguna, lokasi, atau bahkan nama dapat dideteksi oleh model pembelajaran mendalam serupa yang canggih. Untuk mencegah kebocoran privasi, kami akan menambahkan gangguan manipulasi pada gambar asli, sehingga gambar yang dirilis dapat menyesatkan model DNN untuk mendapatkan informasi yang salah. Sementara itu, kami berharap untuk meminimalkan noise sehingga berdampak kecil pada kualitas gambar dan pengalaman pengguna.



Gambar 7.1 Contoh arsitektur sistem untuk perlindungan privasi tingkat file berbasis AP

Kita dapat mendefinisikan masalah perlindungan privasi tingkat file ini sebagai masalah pengoptimalan yang targetnya adalah meminimalkan kemungkinan gambar yang terganggu diklasifikasikan dengan benar oleh penyerang, yaitu,

$$\mathbf{P1}: \min \Pr\left(\boldsymbol{class_P} = \boldsymbol{class_{X|0}}\right),$$

dimana o adalah observasi, classp adalah kelas prediksi dari musuh, dan classX adalah kelas sebenarnya dari gambar asli X.

Output dari P 1 akan berupa angka antara 0 dan 1, dimana "0" berarti sepenuhnya pribadi dan "1" menunjukkan tidak ada privasi. Ada banyak metode berbeda untuk menghasilkan kebisingan untuk contoh manipulasi, di antaranya yang paling banyak digunakan adalah metode tanda gradien cepat (FGSM).

Misalkan θ adalah parameter model, X adalah input model, y adalah target yang terkait dengan X (kita dapat secara acak memilih kelas yang ingin kita menyesatkan model pembelajaran mendalam), dan J(θ X y) menjadi fungsi biaya (output) yang digunakan untuk melatih jaringan saraf. Fungsi biaya dapat dilinierkan di sekitar nilai θ saat ini, sehingga diperoleh gangguan terbatas max-norm optimal

$$\boldsymbol{\eta} = \boldsymbol{\epsilon}\mathbf{sign}(\nabla_{\boldsymbol{X}}\mathbf{J}(\boldsymbol{\theta}; \mathbf{X}; \mathbf{y})),$$

dimana ‹ adalah skalar kecil yang membuat noise tidak terlihat oleh mata manusia dan ∇X adalah gradien fungsi biaya J terhadap gambar masukan X,

$$\nabla_{\boldsymbol{X}}\mathbf{J}(\boldsymbol{\theta}; \mathbf{X}; \mathbf{y}) = \frac{\partial \mathbf{J}}{\partial \mathbf{X}}. .$$

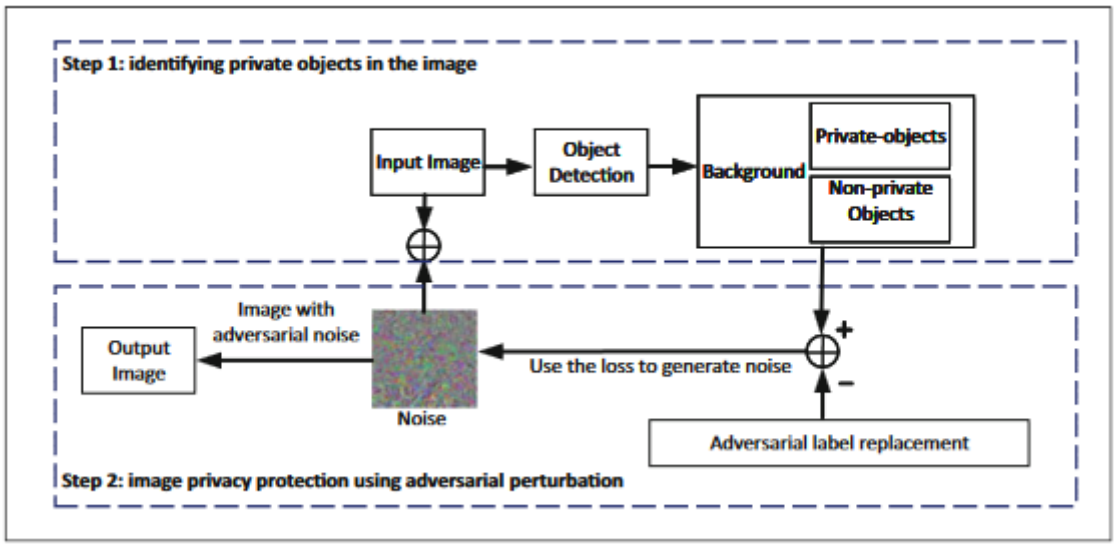
Dan gambar rilis dihasilkan oleh

$$\mathbf{X}' = \boldsymbol{\eta} + \mathbf{X}.$$

Gambar 7.2 memberikan contoh hasil perlindungan privasi tingkat file. Model pembelajaran mendalam memiliki keyakinan yang tinggi (92,42%) untuk mengklasifikasikan gambar asli sebagai "minibus". Dan jika kita menambahkan sedikit noise menggunakan FGSM, maka noise tersebut akan salah diklasifikasikan sebagai "washbasin" dengan tingkat keyakinan yang lebih tinggi (99,37%).

Ketik persamaan di sini.Hasil penelitian yang ada menunjukkan bahwa metode berbasis AP dapat mencapai perlindungan privasi yang baik terhadap alat pembelajaran mendalam dengan mengorbankan sedikit kebisingan yang tidak terlihat oleh mata manusia. Dan efektivitas metode yang diusulkan sangat baik dengan gambar struktur dan tekstur yang kompleks.

Gambar 7.2 Contoh hasil perlindungan privasi tingkat file (warna noise diperkuat dengan normalisasi, jika tidak maka akan sulit dilihat)



Gambar 7.3 Kerangka algoritma perlindungan privasi tingkat objek berbasis AP

Perlindungan Privasi Tingkat Objek

Perlindungan privasi tingkat file cocok untuk gambar sederhana yang hanya berisi satu objek besar. Dalam praktiknya, umumnya terdapat beberapa objek dalam satu gambar, terutama untuk gambar jejaring sosial. Dan beberapa objek sensitif terhadap privasi, sementara objek lainnya mungkin tidak sensitif terhadap privasi. Dalam hal ini, kita dapat menggunakan kerangka perlindungan privasi tingkat objek untuk menyelesaikan masalah.

Seperti ditunjukkan pada Gambar 7.3, kerangka kerja ini dapat terdiri dari dua langkah utama: (i) mengidentifikasi objek pribadi dalam gambar dan (ii) perlindungan privasi gambar menggunakan gangguan konflik. Untuk langkah pertama, dapat digunakan pendeteksi objek berbasis DNN. Jika kita memiliki gambar masukan X, keluaran dari modul deteksi objek direpresentasikan sebagai

$$C(\mathbf{X}) = \left(\begin{array}{cccc|c} x_1 & y_1 & w_1 & h_1 & c_1 \\ x_2 & y_2 & w_2 & h_2 & c_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n & y_n & w_n & h_n & c_n \end{array}\right),$$

$$C(X) = \left(\begin{pmatrix} X_1 & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{pmatrix}\right)$$

dimana xi, yi, wi, dan hi masing-masing mewakili koordinat x pojok kiri atas, koordinat y dan lebar, serta tinggi jangkar. i adalah indeks region of interest (ROI) (i 1, 2,..., n), yang setara dengan jumlah objek pada gambar. cj adalah label kelas (misalnya, kucing, anjing, wajah).

Perlu dicatat bahwa banyak detektor objek seperti Faster RCNN memperlakukan latar belakang sebagai sebuah kelas, yaitu cbg. ambang batas digunakan untuk menangani area yang tidak dapat dikenali yang mungkin muncul. Jika probabilitas semua kelas kurang dari ambang batas, maka hal tersebut dikenali sebagai latar belakang.

Kemudian kita mendefinisikan apa yang dimaksud dengan objek pribadi menurut GDPR:

- Identitas pribadi—plat nomor, nomor telepon, alamat, dll.
- Biometrik—wajah, data kalender, sidik jari, pemindaian retina, foto, dll.
- Catatan elektronik—cookie, lokasi IP, ID perangkat seluler, catatan aktivitas jaringan sosial

Menurut definisi ini, semua kelas dalam keluaran deteksi objek dibagi menjadi dua himpunan bagian: Cprivate adalah himpunan kelas privat, dan Cnon private menyertakan kelas non-privat.

Kemudian pada langkah kedua, gangguan konflik kecil δX yang menargetkan objek pribadi diterapkan untuk menghasilkan gambar bebas privasi Xpr X δX, sehingga hanya informasi non-pribadi yang dapat dideteksi ketika meneruskan Xpr melalui detektor objek, yaitu,

$$C(\mathbf{X}^{pr}) = \left(\begin{array}{cccc|c} x_1 & y_1 & w_1 & h_1 & c_1^{pr} \\ x_2 & y_2 & w_2 & h_2 & c_2^{pr} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n & y_n & w_n & h_n & c_n^{pr} \end{array}\right),$$

Dimana

$$\forall \mathbf{c_j} \in \boldsymbol{C_{private}}: \boldsymbol{C_j^{pr}} = \boldsymbol{C_{bg}}.$$

Berdasarkan kerangka yang dijelaskan di atas, target kami adalah mengelabui jaringan dengan mengubah kelas objek pribadi menjadi latar belakang, sedangkan objek non-pribadi dikenali sebagai kelas aslinya. Sedangkan kebisingan tambahan δX harus kecil agar tidak terlihat oleh manusia. Oleh karena itu permasalahannya dapat dirumuskan sebagai berikut:

$$\arg\min_{\delta X} \|\delta X\|_2$$

$$\text{s.t.: } \forall c_j \in \mathbf{C}_{private} : c_j^{pr} = c_{bg}$$

$$\forall c_j \in \mathbf{C}_{non\text{-}private} : c_j^{pr} = c_j$$

Algoritma perlindungan privasi gambar berbasis AP dapat digunakan untuk mengatasi masalah di atas. Seperti yang ditunjukkan pada Gambar 7.3, detektor objek menemukan semua objek dalam gambar di awal. Kemudian, kita mengganti label objek pribadi dengan latar belakang dan menggunakan fungsi kerugian yang sesuai untuk menghitung gradien. Kemudian noise diperbarui sesuai dengan gradiennya. Terakhir, gambar yang terganggu dihasilkan, di mana semua objek privasi diperlakukan sebagai latar belakang oleh detektor objek.

Bagian penting dari algoritme ini adalah mengelabui kerugian klasifikasi (cls) sehingga menyesatkan detektor objek yang mengenali objek privasi di latar belakang, seperti yang ditunjukkan pada Persamaan. (7.1):

$$\mathbf{L}_{CLS} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}} \boldsymbol{En}\,(\boldsymbol{P_i}, \boldsymbol{P_i}) + \boldsymbol{\lambda} \,||\, \mathbf{X} - \mathbf{X}^{\boldsymbol{pr}} \,||_{\mathbf{2}} \qquad (\mathbf{7,1})$$

dimana pi pi$_1$,..., pim adalah probabilitas konten sebuah jangkar dikenali oleh setiap kelas. pi dikodekan one-hot (pi∗ 0, 0,..., 1,..., 0, 0), di mana 1 muncul di posisi di mana kita menetapkan kelas sebagai kelas yang benar. pi∗ akan dihasilkan sesuai dengan label kebenaran dasar jika objek tersebut non-pribadi, sedangkan akan diubah ke latar belakang jika objek tersebut bersifat pribadi. n adalah jumlah total objek dalam gambar sehingga entropi akan dirata-ratakan pada semua jangkar. Selanjutnya, kita dapat menggunakan cls untuk menghasilkan gangguan, menggunakan metode tanda gradien cepat (FGSM).

Dengan menggunakan FGSM yang ditargetkan, gangguan dapat dihitung berdasarkan arah gradien:

$$\boldsymbol{\delta \mathbf{X}} = -\in \mathbf{sign}\,(\nabla_{\boldsymbol{X}} \boldsymbol{L_{cls}}) = -\in \boldsymbol{sign}\left(\frac{\boldsymbol{L_{cls}}}{\boldsymbol{\delta \mathbf{X}}}\right),$$

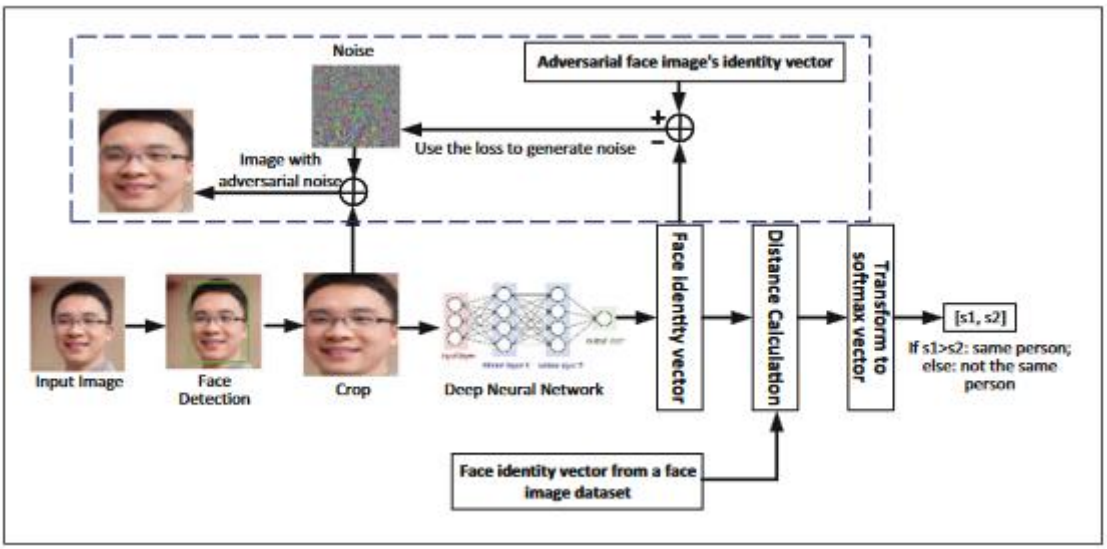
dimana ‹ adalah parameter langkah yang menskalakan kebisingan. Oleh karena itu, gambar yang dihasilkan adalah:

$$\mathbf{X^{pr}} = \mathbf{X} + \boldsymbol{\delta \mathbf{X}} = \mathbf{X} - \in \mathbf{sign}\,(\frac{\mathbf{L_{cls}}}{\boldsymbol{\delta \mathbf{X}}})$$

Perlindungan Privasi Tingkat Fitur

Dalam beberapa kasus lain, kita hanya perlu mengubah fitur tertentu dalam gambar atau video, menggunakan gangguan manipulasi yang tidak terlihat oleh manusia. Contoh tipikalnya adalah mengubah identitas seseorang (terhadap sistem pengenalan wajah) pada gambar sambil menjaga tampilan visualnya tidak berubah.

Sistem pengenalan wajah merupakan teknologi yang mampu mengenali atau mengautentikasi seseorang dari suatu gambar atau bingkai video. Dengan jaringan saraf pembelajaran mendalam yang canggih baru-baru ini, keakuratan sistem pengenalan wajah berbasis kecerdasan buatan mulai melampaui akurasi manusia dalam beberapa pengujian benchmark. Hasilnya, mereka mulai melihat kegunaan yang lebih luas di banyak aplikasi, seperti kontrol akses dan pemantauan keamanan.



Gambar 7.4 Ilustrasi sistem pengenalan wajah pada umumnya dan proses menghasilkan gangguan gambar yang berlawanan

Gambar 7.4 menggambarkan sistem pengenalan wajah pada umumnya. Saat gambar masukan diterima, pertama-tama ia mendeteksi posisi wajah dan memotong wajah ke ukuran yang sesuai dengan pengaturan sistem. DNN digunakan untuk menghitung penyematan wajah (vektor numerik yang mewakili fitur wajah) dari gambar wajah. Kemudian sistem dapat menghitung jarak antara penyematan permukaan masukan dan penyematan tertentu dari database sistem. Jarak diubah menjadi vektor yang berisi dua nilai lunak yang menunjukkan hasil pengenalan wajah: jika nilai pertama lebih besar dari nilai kedua, maka kedua penyematan tersebut berasal dari gambar orang yang sama. Kalau tidak, itu adalah gambaran dari dua orang yang berbeda.

Dalam arti tertentu, sistem pengenalan wajah mirip dengan seseorang dalam melakukan tugas mengenali orang lain: orang tersebut membandingkan gambar baru dengan ingatannya. Jika gambar tersebut terlihat dekat dengan seseorang dalam ingatannya, mereka

menganggapnya sebagai orang yang sama. Satu-satunya perbedaan adalah cara DNN dan manusia mengukur "jarak" antar gambar.

Dari perspektif perlindungan privasi, kami bertujuan untuk menambahkan noise pada gambar asli sehingga sistem pengenalan wajah tidak dapat mengidentifikasi orang tersebut dengan benar. Secara lebih rinci, berdasarkan metrik tingkat keberhasilan perlindungan privasi, usulan masalah perlindungan privasi gambar dapat dirumuskan sebagai:

$$\boldsymbol{P2}: \boldsymbol{max}\ \boldsymbol{Pr}(\boldsymbol{I}\ \boldsymbol{D}_{\boldsymbol{X}'} \neq \boldsymbol{I}\ \boldsymbol{D}_{\boldsymbol{X}}), \qquad (\boldsymbol{7}, \boldsymbol{2})$$

dimana IDX adalah identitas gambar asli dan IDXr adalah identitas gambar yang mengalami gangguan. Gangguan manipulasi dapat dihasilkan dengan algoritma FGSM atau metode yang lebih kuat, yaitu varian multistep FGSMN, yang pada dasarnya merupakan proyeksi penurunan gradien (PGD) pada fungsi kerugian negatif.

Dalam PGD, FGSM akan diulang sebanyak N kali atau hingga nilai absolut kebisingan mencapai batas atas yang telah ditentukan, yaitu,

$$\begin{aligned} \boldsymbol{X}'_{\boldsymbol{0}} &= \boldsymbol{X} \\ \boldsymbol{X}'_{\boldsymbol{n}} &= \boldsymbol{X}'_{\boldsymbol{n-1}} + \in \boldsymbol{sign}\left(\nabla_{\boldsymbol{x}}\, \boldsymbol{j}(\,\boldsymbol{\theta};\, \boldsymbol{X}'_{\boldsymbol{n-1}}; \boldsymbol{y})\right) \\ &= \boldsymbol{X}'_{\boldsymbol{n-1}} + \boldsymbol{\eta}_{\boldsymbol{n-1'}} \boldsymbol{1} \leq \boldsymbol{n} \leq \boldsymbol{N}. \end{aligned}$$

Ilustrasi proses ditunjukkan pada Gambar 7.4. Pertama, orang yang berbeda dipilih secara spesifik atau acak. Kemudian vektor penyematan permukaan adversarial ini akan dihitung dan dijadikan nilai y. Citra dengan gangguan adversarial dihasilkan oleh algoritma PGD dan terakhir diuji menggunakan sistem pengenalan wajah.

7.3 DISKUSI DAN PEKERJAAN MASA DEPAN

Meskipun metode berbasis AP telah menunjukkan efektivitas perlindungan privasi yang luar biasa bahkan pada tingkat kebisingan yang tidak terlihat, saat ini ada dua masalah besar dengan kelompok metode ini: (1) metode ini sangat bergantung pada aksesibilitas ke sistem target, sehingga hanya dapat dijamin untuk pengenalan target spesifik (yaitu, memerlukan pengetahuan kotak putih), dan (2) kemampuan transfer gangguan manipulasi, yaitu efektivitasnya pada model alternatif yang tidak diketahui tidak sebaik terhadap model target.

Untuk mengatasi masalah di atas, beberapa makalah telah mentransfer perhitungan arah kebisingan dari lapisan keluaran ke lapisan perantara model. Hal ini dapat menghindari perbedaan antar model, sehingga meningkatkan kemampuan transfer. Makalah Pidhorskyi mempelajari potensi penambahan gangguan manipulasi pada tingkat fitur gambar. Karena model DNN yang berbeda memiliki keluaran yang serupa dalam tingkat fitur, hal ini juga akan meningkatkan kemampuan transfer.

Dari perspektif perlindungan privasi, ada beberapa mekanisme lainnya. Misalnya, ada beberapa peneliti yang mulai menggunakan GAN untuk menghasilkan konten guna menggantikan informasi sensitif dalam gambar. Matahari dkk. mengusulkan pengecatan

kepala berbasis GAN untuk menghilangkan identitas aslinya. Selain itu, baru-baru ini ada beberapa upaya untuk menggabungkan gagasan DP dengan privasi gambar. Fan mengusulkan metode privat ‹-diferensial pada tingkat piksel gambar. Namun, membuat piksel gambar tidak dapat dibedakan dalam praktiknya tidak masuk akal, dan kualitas gambar yang dihasilkan cukup rendah. Ini akan menjadi topik yang menarik untuk membandingkan mekanisme perlindungan privasi yang berbeda. Terakhir, setelah tahap pertama perlindungan privasi gambar, penelitian tentang perlindungan privasi video juga telah dimulai. Karena penerapan langsung metode perlindungan privasi gambar yang ada pada video akan menimbulkan kompleksitas komputasi yang tinggi dan latensi yang besar, merancang mekanisme perlindungan privasi video yang lebih efektif juga merupakan arah penelitian yang menjanjikan.

DAFTAR PUSTAKA

Adesina, D., Hsieh, C. C., Sagduyu, Y. E., & Qian, L. (2022). Adversarial machine learning in wireless communications using RF data: A review. *IEEE Communications Surveys & Tutorials*, *25*(1), 77-100.

Albert, K., Penney, J., Schneier, B., & Kumar, R. S. S. (2020). Politics of adversarial machine learning. *arXiv preprint arXiv:2002.05648*.

Alhajjar, E., Maxwell, P., & Bastian, N. (2021). Adversarial machine learning in network intrusion detection systems. *Expert Systems with Applications*, *186*, 115782.

Alsmadi, I., Aljaafari, N., Nazzal, M., Alhamed, S., Sawalmeh, A. H., Vizcarra, C. P., ... & Al-Humam, A. (2022). Adversarial machine learning in text processing: a literature survey. *IEEE Access*, *10*, 17043-17077.

Apruzzese, G., Colajanni, M., Ferretti, L., & Marchetti, M. (2019, May). Addressing adversarial attacks against security systems based on machine learning. In *2019 11th international conference on cyber conflict (CyCon)* (Vol. 900, pp. 1-18). IEEE.

Biggio, B., & Roli, F. (2018, October). Wild patterns: Ten years after the rise of adversarial machine learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security* (pp. 2154-2156).

Biggio, B., Russu, P., Didaci, L., & Roli, F. (2015). Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective. *IEEE Signal Processing Magazine*, *32*(5), 31-41.

Brendel, W., Rauber, J., & Bethge, M. (2017). Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*.

Bulò, S. R., Biggio, B., Pillai, I., Pelillo, M., & Roli, F. (2016). Randomized prediction games for adversarial machine learning. *IEEE transactions on neural networks and learning systems*, *28*(11), 2466-2478.

Chen, P. Y., & Hsieh, C. J. (2022). *Adversarial robustness for machine learning*. Academic Press.

Chen, S., Xue, M., Fan, L., Hao, S., Xu, L., Zhu, H., & Li, B. (2018). Automated poisoning attacks and defenses in malware detection systems: An adversarial machine learning approach. *computers & security*, *73*, 326-344.

De Lucia, M. J., & Cotton, C. (2019). Adversarial machine learning for cyber security. *Journal of Information Systems Applied Research*, *12*(1), 26.

Deldjoo, Y., Di Noia, T., & Merra, F. A. (2020, January). Adversarial machine learning in recommender systems (aml-recsys). In *Proceedings of the 13th International Conference on Web Search and Data Mining* (pp. 869-872).

Delobelle, P., Temple, P., Perrouin, G., Frénay, B., Heymans, P., & Berendt, B. (2021). Ethical adversaries: Towards mitigating unfairness with adversarial machine learning. *ACM SIGKDD Explorations Newsletter*, *23*(1), 32-41.

Duddu, V. (2018). A survey of adversarial machine learning in cyber warfare. *Defence Science Journal*, *68*(4), 356.

Edwards, D., & Rawat, D. B. (2020). Study of adversarial machine learning with infrared examples for surveillance applications. *Electronics*, *9*(8), 1284.

Edwards, D., & Rawat, D. B. (2020, October). Quantum adversarial machine learning: Status, challenges and perspectives. In *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)* (pp. 128-133). IEEE.

Englert, C., Galler, P., Harris, P., & Spannowsky, M. (2019). Machine learning uncertainties with adversarial neural networks. *The European Physical Journal C*, *79*, 1-10.

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science*, *363*(6433), 1287-1289.

He, K., Kim, D. D., & Asghar, M. R. (2023). Adversarial machine learning for network intrusion detection systems: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, *25*(1), 538-566.

Hernández-Castro, C. J., Liu, Z., Serban, A., Tsingenopoulos, I., & Joosen, W. (2022). Adversarial machine learning. In *Security and Artificial Intelligence: A Crossdisciplinary Approach* (pp. 287-312). Cham: Springer International Publishing.

Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I., & Tygar, J. D. (2011, October). Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence* (pp. 43-58).

Jang, U., Wu, X., & Jha, S. (2017, December). Objective metrics and gradient descent algorithms for adversarial examples in machine learning. In *Proceedings of the 33rd Annual Computer Security Applications Conference* (pp. 262-277).

Jmila, H., & Khedher, M. I. (2022). Adversarial machine learning for network intrusion detection: A comparative study. *Computer Networks*, *214*, 109073.

Joseph, A. D., Nelson, B., Rubinstein, B. I., & Tygar, J. D. (2018). *Adversarial machine learning*. Cambridge University Press.

Khachaturov, D., Shumailov, I., Zhao, Y., Papernot, N., & Anderson, R. (2021, July). Markpainting: Adversarial machine learning meets inpainting. In *International Conference on Machine Learning* (pp. 5409-5419). PMLR.

Khan, M., & Ghafoor, L. (2024). Adversarial Machine Learning in the Context of Network Security: Challenges and Solutions. *Journal of Computational Intelligence and Robotics*, *4*(1), 51-63.

Kumar, A., Mehta, S., & Vijaykeerthy, D. (2017, November). An introduction to adversarial machine learning. In *International Conference on Big Data Analytics* (pp. 293-299). Cham: Springer International Publishing.

Kumar, R. S. S., Nyström, M., Lambert, J., Marshall, A., Goertzel, M., Comissoneru, A., ... & Xia, S. (2020, May). Adversarial machine learning-industry perspectives. In *2020 IEEE security and privacy workshops (SPW)* (pp. 69-75). IEEE.

Kumar, R. S. S., O'Brien, D. R., Albert, K., & Vilojen, S. (2018). Law and adversarial machine learning. *arXiv preprint arXiv:1810.10731*.

Kumar, R. S. S., Penney, J., Schneier, B., & Albert, K. (2020). Legal risks of adversarial machine learning research. *arXiv preprint arXiv:2006.16179*.

Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial examples in the physical world. In *Artificial intelligence safety and security* (pp. 99-112). Chapman and Hall/CRC.

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Laskov, P., & Lippmann, R. (2010). Machine learning in adversarial environments. *Machine learning*, *81*, 115-119.

Li, G., Zhu, P., Li, J., Yang, Z., Cao, N., & Chen, Z. (2018). Security matters: A survey on adversarial machine learning. *arXiv preprint arXiv:1810.07339*.

Li, J., Yang, Y., Sun, J. S., Tomsovic, K., & Qi, H. (2021, May). Conaml: Constrained adversarial machine learning for cyber-physical systems. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security* (pp. 52-66).

Lin, H. Y., & Biggio, B. (2021). Adversarial machine learning: Attacks from laboratories to the real world. *Computer*, *54*(5), 56-60.

Liu, J., Nogueira, M., Fernandes, J., & Kantarci, B. (2021). Adversarial machine learning: A multilayer review of the state-of-the-art and challenges for wireless and mobile systems. *IEEE Communications Surveys & Tutorials*, *24*(1), 123-159.

Lu, S., Duan, L. M., & Deng, D. L. (2020). Quantum adversarial machine learning. *Physical Review Research*, *2*(3), 033212.

Luo, Z., Zhao, S., Lu, Z., Sagduyu, Y. E., & Xu, J. (2020, July). Adversarial machine learning based partial-model attack in IoT. In *Proceedings of the 2nd ACM workshop on wireless security and machine learning* (pp. 13-18).

Machado, G. R., Silva, E., & Goldschmidt, R. R. (2021). Adversarial machine learning in image classification: A survey toward the defender's perspective. *ACM Computing Surveys (CSUR)*, *55*(1), 1-38.

Martins, N., Cruz, J. M., Cruz, T., & Abreu, P. H. (2020). Adversarial machine learning applied to intrusion and malware scenarios: a systematic review. *IEEE Access*, *8*, 35403-35419.

McDaniel, P., Papernot, N., & Celik, Z. B. (2016). Machine learning in adversarial settings. *IEEE Security & Privacy*, *14*(3), 68-72.

Na, T., Ko, J. H., & Mukhopadhyay, S. (2017). Cascade adversarial machine learning regularized with a unified embedding. *arXiv preprint arXiv:1708.02582*.

Nowroozi, E., Dehghantanha, A., Parizi, R. M., & Choo, K. K. R. (2021). A survey of machine learning techniques in adversarial image forensics. *Computers & Security*, *100*, 102092.

Pacheco, Y., & Sun, W. (2021, February). Adversarial Machine Learning: A Comparative Study on Contemporary Intrusion Detection Datasets. In *ICISSP* (pp. 160-171).

Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*.

Rawal, A., Rawat, D., & Sadler, B. M. (2021). Recent advances in adversarial machine learning: status, challenges and perspectives. *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III*, *11746*, 701-712.

Rosenberg, I., Shabtai, A., Elovici, Y., & Rokach, L. (2021). Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*, *54*(5), 1-36.

Sadeghi, K., Banerjee, A., & Gupta, S. K. (2020). A system-driven taxonomy of attacks and defenses in adversarial machine learning. *IEEE transactions on emerging topics in computational intelligence*, *4*(4), 450-467.

Tabassi, E., Burns, K. J., Hadjimichael, M., Molina-Markham, A. D., & Sexton, J. T. (2019). A taxonomy and terminology of adversarial machine learning. *NIST IR*, *2019*, 1-29.

Taheri, R., Javidan, R., Shojafar, M., Vinod, P., & Conti, M. (2020). Can machine learning model with static features be fooled: an adversarial machine learning approach. *Cluster computing*, *23*, 3233-3253.

Tramèr, F., Dupré, P., Rusak, G., Pellegrino, G., & Boneh, D. (2019, November). Adversarial: Perceptual ad blocking meets adversarial machine learning. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security* (pp. 2005-2021).

Tygar, J. D. (2011). Adversarial machine learning. *IEEE Internet Computing*, *15*(5), 4-6.

Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2024). Adversarial machine learning. *Gaithersburg, MD*.

Vorobeychik, Y., Kantarcioglu, M., Brachman, R., Stone, P., & Rossi, F. (2018). Adversarial machine learning.

Wang, X., Li, J., Kuang, X., Tan, Y. A., & Li, J. (2019). The security of machine learning in an adversarial setting: A survey. *Journal of Parallel and Distributed Computing*, *130*, 12-23.

West, M. T., Tsang, S. L., Low, J. S., Hill, C. D., Leckie, C., Hollenberg, L. C., ... & Usman, M. (2023). Towards quantum enhanced adversarial robustness in machine learning. *Nature Machine Intelligence*, *5*(6), 581-589.

Wiyatno, R. R., Xu, A., Dia, O., & De Berker, A. (2019). Adversarial examples in modern machine learning: A review. *arXiv preprint arXiv:1911.05268*.

Xi, B. (2020). Adversarial machine learning for cybersecurity and computer vision: Current developments and challenges. *Wiley Interdisciplinary Reviews: Computational Statistics*, *12*(5), e1511.

Zhou, Y., Kantarcioglu, M., & Xi, B. (2019). A survey of game theoretic approach for adversarial machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), e1259.

Zhou, Y., Kantarcioglu, M., Thuraisingham, B., & Xi, B. (2012, August). Adversarial support vector machine learning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1059-1067).

Zizzo, G., Hankin, C., Maffeis, S., & Jones, K. (2019, June). Adversarial machine learning beyond the image domain. In *Proceedings of the 56th Annual Design Automation Conference 2019* (pp. 1-4).

CARA MEMANIPULASI PEMBELAJARAN MESIN

(MACHINE LEARNING)

Dr. Joseph Teguh Santoso, S.Kom, M.Kom

BIODATA PENULIS

Dr. Joseph Teguh Santoso, S.Kom, M.Kom adalah Rektor dari Universitas Sains & Teknologi Komputer (Universitas STEKOM) Semarang yang memiliki banyak pengalaman praktis dalam bidang *e-commerce*sejak Tahun 2002. Beliau mempunyai 3 (tiga) toko *Official Online Store* di China untuk merek Sepeda Raleigh, dengan omzet tahunan pada Tahun 2019 mencapai lebih dari Rp. 35 Milyar rupiah dan terus meningkat. Dr. Joseph T.S memiliki lisensi tunggal sepeda merek "Raleigh" untuk penjualan *Online* di seluruh China. Di samping itu beliau juga memiliki pabrik sepeda dan sepeda listrik merek "Fengjiu", yaitu Pabrik Sepeda Listrik yang masih tergolong kecil di China. Pengalaman beliau malang melintang di dunia *online store* di China seperti Alibaba, Tmall, Taobao, JD, Aliexpress sangat membantu mahasiswa untuk memiliki pengalaman teknis dan praktis untuk membuka toko *online* bersama beliau.

YAYASAN PRIMA AGUS TEKNIK

PENERBIT :
YAYASAN PRIMA AGUS TEKNIK
Jl. Majapahit No. 605 Semarang
Telp. (024) 6723456. Fax. 024-6710144
Email : penerbit_ypat@stekom.ac.id

ISBN 978-623-8642-00-7 (PDF)