ارائهٔ یک روش بر پایهٔ ماشین‌های بردار پشتیبان برای تشخیص رخنه به شبکه‌های رایانه‌ای

نرگس صالح پور[۱]، محمد نظری فرخی[۲] و ابراهیم نظری فرخی[۳]

[۱و۲] کارشناسی ارشد کامپیوتر (نرم‌افزار) ، دانشگاه آزاد اسلامی واحد علوم تحقیقات لرستان، لرستان، ایران.

Salehpour_Narges@yahoo.com

M_Kasitman@yahoo.com

[۳] دانشجوی دکتری مدیریت فناوری اطلاعات، دانشگاه آزاد اسلامی واحد علوم تحقیقات، تهران، ایران.

[3]E60_Itmgtn@yahoo.com

چکیده

سامانهٔ تشخیص رخنه یکی از مهم‌ترین مسائل در تأمین امنیت شبکه‌های رایانه‌ای است. سامانه‌های تشخیص رخنه در جستجوی رفتار مخرب، انحراف الگوهای طبیعی و کشف حملات به شبکه‌های رایانه‌ای هستند. این سامانه‌ها نوع ترافیک مجاز از ترافیک غیرمجاز را تشخیص می‌دهند. از آن‌جا که امروزه تکنیک‌های داده‌کاوی به‌منظور تشخیص رخنه در شبکه‌های رایانه‌ای مورد استفاده قرار می‌گیرند، در این پژوهش نیز، روشی مبتنی‌بر یادگیری ماشین جهت طراحی یک سامانه تشخیص رخنه ارائه شده است. یکی از ویژگی‌های شبکه‌های عصبی و سامانه‌های یادگیری ماشین، آموزش بر اساس داده‌های آموزشی است. در این پژوهش برای تشخیص رخنه از یادگیری ماشین با خاصیت یادگیری روی ویژگی‌ها با استفاده از تئوری راف که دارای ضریب همبستگی بیشتری است، به‌کار گرفته می‌شود. برای آموزش و ارزیابی روش پیشنهادی از مجموعه دادهٔ KDD CUP 99 استفاده شده است. بنابراین دقت روش پیشنهادی را با الگوریتم یادگیری بر پایهٔ تمام ویژگی‌ها، شبکه عصبی خودسازمانده و درخت تصمیم‌گیری مقایسه می‌کند. نتایج شبیه‌سازی نشان می‌دهد، سامانه پیشنهادی مبتنی بر تئوری راف دارای دقت بالا و سرعت تشخیص مناسب است.

واژگان کلیدی: تئوری مجموعه‌های راف، سامانه‌های تشخیص رخنه، ماشین بردار پشتیبان و سامانه‌های یادگیری ماشین.

۱- مقدّمه

سامانهٔ تشخیص رخنه در مورد چگونگی حمله و یا این‌که چه کسی سعی در رخنه به سامانه را دارد، اطلاعاتی در اختیار کاربر قرار می‌دهد. با توجه به بررسی‌های انجام‌شده در مرکز سی. إس. آی[۱]، در دوازدهمین کنگره سالانه جرایم رایانه‌ای و بررسی امنیت که با حضور بیش از ۴۰۰ شرکت پیشگام در زمینه امنیت سامانه‌های رایانه‌ای در مؤسسات، نهادهای حکومتی، دانشگاه و... در ایالت متحده آمریکا انجام شده بود، می‌توان دریافت که با افزایش سرمایه‌گذاری بر روی صنعت سامانه‌های رایانه‌ای، میزان حملات به‌طور چشم‌گیری کاهش یافته است. با توجه به آمار، از سال ۲۰۰۰ تا ۲۰۰۷ استفاده‌کنندگان از اینترنت به بیش از یک میلیارد می‌رسید.

[1] CSI

با وجود اینکه، امروزه تعداد کاربران اینترنت تا ۲۱۴ درصد رشد داشته است و با رشد کاربران اینترنت، رخنه به سامانه‌های رایانه‌ای از طریق شبکه نیز افزایش یافته است، بنابراین امنیت شبکه به یک مسئله جدی برای کاربران شخصی و شرکت‌های بزرگ تبدیل شده است. به‌طورکلی امنیت، شامل ضدهرزنامه، ضدویروس، ورود و خروج حسابرسی‌ها و دیوارهای‌آتش است که ضدهرزنامه بر روی کاربران نهایی تمرکز دارد و رایانامه‌هایی را که در هرزنامه قرار می‌گیرند، مسدود می‌کند. تکنیک‌های ضدویروس نیز به‌طورمعمول برای شناسایی نرم‌افزار، بلوک و از بین‌بردن ویروس‌های رایانه‌ای و برنامه‌های مخرب استفاده می‌شوند. یکی دیگر از راه‌های معمول برای رخنه پیداکردن به شبکه‌های رایانه‌ای، استفاده از نام کاربری و رمز عبور است که

در سال ۱۹۸۰ توسط Denning شناخته شد. ازجمله سامانه‌های تشخیص تجاری تریپ وایر[1]، امنیت مانیتور کین[2]، پلیس سایبر[3]، امنیت واقعی[4]، رنجر شبکه[5]، سنتراکس[6]، امنیت سامانه‌های تشخیص رخنه سیسکو[7]، اژدها[8]، اینترورت[9] را می‌توان نام برد. برای اطلاعات بیشتر می‌توان به مراجع [۱،٤،۳،۲] مراجعه کرد. در حال حاضر سیستم‌عامل‌ها و شبکه‌های رایانه‌ای از حس‌گرهای تشخیص رخنه برای شناسایی حملات و رخنه‌گران استفاده می‌کنند. برخی، پاسخ به رخنه‌های غیرقانونی سامانه را پایان‌دادن اتصال به شبکه در نظر می‌گیرند. نمونه‌هایی از سطح رخنه در شبکه شامل محرومیت از راه دور، ورود تروجان‌ها و جاسوس‌ها به سامانه را می‌توان نام برد[٦، ٥]. هدف اصلی، ارائهٔ یک روش مبتنی بر الگوریتم‌های یادگیری ماشین ازجمله ماشین‌های بردار پشتیبان، رگرسیون ماشین بردار پشتیبان در تشخیص رخنه است که در این پژوهش به‌عنوان یادگیری ماشین از آن‌ها استفاده شده است. روش ذکرشده به‌منظور به‌کارگیری روش‌های تشخیص، جلوگیری از نفوذها ازجمله حملات سایبری، ویروس‌ها، کرم‌های رایانه‌ای و غیره است. با بررسی الگوی فعالیت کاربران با استفاده از شبکه‌های عصبی و ماشین‌های بردار پشتیبان بسیاری از نفوذها قابل تشخیص هستند. الگوهای ذخیره‌شده باید به‌طور مداوم به‌روز شوند که این امر مستلزم تخصص و ابزار هوشمند است. ماشین بردار پشتیبان یک ابزار عملی برای مجموعه‌داده‌های بزرگ است که داده‌ها را با تعیین مجموعه‌داده‌های بردار پشتیبان دسته‌بندی می‌کند. هم‌چنین خطای تعمیم را نیز به حداقل می‌رسانند. اگرچه ماشین بردار پشتیبان روش امیدوارکننده‌ای برای دسته‌بندی داده‌ها است، اما مشکل اصلی آن پیچیدگی آموزش‌های بسیار، وابسته به مجموعه‌داده‌ها است. برای این منظور از ترکیب تکنیک دیگری به نام تئوری مجموعه‌های راف با الگوریتم‌های یادگیری ماشین برای شناسایی حملات استفاده می‌شود.

۲- سامانه‌های تشخیص رخنه

با گسترش روزافزون تبادل اطلاعات و سامانه‌های برخط، میزان حملات و رخنه در سامانه‌های اطلاعاتی، افزایش یافته است. به‌منظور جلوگیری از سوءاستفادهٔ اطلاعات، امنیت شبکه باید در نظر گرفته شود. مفهوم امنیت به معنای محافظت از:

- **محرمانگی**: محرمانه‌بودن داده‌ها به معنی آن است که داده‌های در حال انتقال در شبکه تنها باید توسط افرادی که به شیوهٔ مناسب، احراز هویت شده‌اند، مورد دستیابی قرار گیرند و داده‌ها از لحظه‌ای که منتقل تا لحظه‌ای که دریافت می‌شوند، نباید هیچ اختلال و فقدانی به‌دلیل فعالیت‌های خراب‌کارانه داشته باشند.
- **دست‌نخوردگی**: اطلاعات ممکن است توسط مهاجم تغییریافته یا نابود شوند.
- **دسترس‌پذیری**: دسترسی‌پذیری در حالت مشهود به معنای این است که سامانه در مقابل حملات انکار سرویس (DOS)[10]، محافظ باشد [۳۴، ۲۱].

تشخیص رخنه عبارت است از فرایند شناسایی و پاسخ به فعالیت‌های مخرب که به‌صورت هدفمند، منابع شبکه را مورد تهاجم قرار می‌دهند. اگرچه سطوح بسیاری وجود دارند که از دسترسی به شبکه‌های رایانه‌ای محافظت می‌کنند؛ اما رخنه‌گران راه‌هایی را برای ورود به شبکه می‌یابند تا خسارات عمده‌ای را در سطح شبکه به وجود آورند. هدف سامانه‌های تشخیص رخنه، جلوگیری از حمله نیست؛ بلکه کشف و شناسایی حملات و تشخیص اشکالات امنیتی در شبکه‌های رایانه‌ای و درنهایت، اعلام آن به مدیران سامانه است.

ازجمله وظایف سامانه‌های تشخیص رخنه عبارتنداز [۲۲]:

- تأیید خطاهای سامانه؛
- بررسی یک‌پارچگی سامانه و پرونده‌های داده؛
- تشخیص رفتارهای غیرعادی و نابه‌هنجار سامانه؛
- شناسایی حملات و هشدارها.

سامانهٔ هشدار چندگانه[11] یکی از سامانه‌های تشخیص رخنه است که به‌عنوان سامانهٔ خبره شناخته شده است و برای تجزیه‌وتحلیل داده‌ها مورد استفاده قرار می‌گیرد. یکی دیگر از سامانه‌های تشخیص رخنه، NIDES[12] است که به‌عنوان سامانهٔ تشخیص رخنهٔ جامع مطرح می‌شود و برای تشخیص ناهنجاری و سوءاستفادهٔ پیاده‌سازی‌شده است. NIDES، ناهنجاری را با استفاده از پروفایل (نمایه)

[1] Tripwire
[2] Kane Security monitor
[3] Cyber cop
[4] Real Secure
[5] NetRanger
[6] Centrax
[7] Cisco Secure IDS
[8] Dragon
[9] Intruvert

[10] Denial of Service Attacks
[11] Multics Intrusion Detection and Alerting System
[12] Next generation intrusion detection expert system

تشخیص می‌دهد. نمایه‌ها درواقع الگوهای عادی فعالیت‌های سامانه را ارائه می‌دهند. به‌طورمعمول پروفایل‌ها در NIDES یک‌بار در روز تغییرات جدید را به‌روز می‌کند.

جیمز اندرسون رخنه را به دو دسته رخنهٔ داخلی و خارجی تقسیم کرد. آقای اندرسون بر روی مجموعه‌ای از رکوردهایی که رفتار غیرمعمول سامانه را بیان می‌کرد، مانند استفاده از زمان غیرمجاز، دفعات استفاده غیرمجاز و الگوی غیرمعمول از ارجاع به برنامه‌ها و داده‌ها، متمرکز شد. وی همچنین در مورد مشکلات دسترسی کاربران مجاز به اطلاعات حیاتی سامانه، نیز هشدار داد و معلوم شد که تشخیص استفاده غیرمجاز از رکوردهای امنیتی، کار سختی است[۲۳].

۲-۱- انواع سامانه‌های تشخیص رخنه از نظر واکنش

سامانه‌های تشخیص رخنه از نظر نوع واکنش به دو دسته تقسیم می‌شوند:

- **سامانهٔ تشخیص رخنه مبتنی بر انفعال**

با توجه به این که سامانه‌های تشخیص نفوذ برای کمک به مدیران امنیتی سامانه در جهت کشف رخنه و حمله به کار گرفته می‌شوند، بنابراین این سامانه‌ها فقط حملات را شناسایی و پیغام لازم را به مدیر اعلام می‌کنند.

- **سامانهٔ تشخیص رخنه مبتنی برعکس العمل[1]**

این سامانه‌ها علاوه‌بر شناسایی حملات به مدیر سامانه پیغام می‌دهند. همچنین راه ارتباطی فرستنده مورد نظر را نیز به شبکه مسدود می‌کنند و نمی‌گذارند هیچ بسته‌ای از طرف IP[2] به شبکه وارد شود[۳۷].

۲-۲- سامانه‌های تشخیص رخنه از نظر نوع تشخیص

روش‌های تشخیص سوءاستفاده برای به رسمیت شناختن الگوی حملات شناخته می‌شوند. ازجمله روش‌های تشخیص رخنه، از نظر نوع تشخیص به دو دسته تقسیم می‌شوند:

- **تشخیص مبتنی بر امضا[3]**

در این روش الگوهای رخنه از پیش‌ساخته‌شده و به‌صورت قانونی نگه‌داری می‌شوند. تشخیص‌دهنده دارای پایگاه داده‌ای از امضاها یا الگوهای حمله است. البته این روش قادر به تشخیص نفوذهای شناخته‌شده است؛ و در صورت بروز حملهٔ جدید در سطح شبکه قادر به شناسایی نیست. تشخیص بر اساس امضا، ساده‌ترین روش است؛ زیرا در شناسایی تهدیداتی که از قبل شناخته‌شده باشند، بسیار مؤثر است و در مورد شناسایی حملاتی که از قبل شناخته‌شده نیستند، به‌طورکامل غیر مؤثر است. یکی از معروف‌ترین امضاهای مبتنی بر سامانه‌های تشخیص رخنه، کامپلیتلی[4] است.

- **تشخیص مبتنی بر ناهنجاری**

تشخیص مبتنی بر ناهنجاری عبارت است از مقایسهٔ رویدادهای مشاهده‌شده با فعالیت‌های نرمال، تا میزان تفاوت آن‌ها مشخص شود. برای تشخیص ناهنجاری باید الگوهای خاصی را پیدا کرد و رفتارهایی که از الگو پیروی می‌کنند، عادی و رویدادهای انحرافی به‌عنوان ناهنجاری تشخیص داده می‌شوند. ویژگی اصلی تشخیص مبتنی بر ناهنجاری این است که برای حملات ناشناخته و جدید نیز مؤثر است [۳۲].

۲-۲- سامانه‌های تشخیص رخنه، بر اساس منابع دریافتی اطلاعات

به‌منظور شناسایی و تشخیص فعالیت‌های غیرمجاز رایانه براساس منابع دریافت اطلاعات سامانه‌های تشخیص رخنه به دو دسته تقسیم می‌شوند:

- **سامانه‌های مبتنی بر میزبان[5]**

در این سامانه‌ها اطلاعات از چندین میزبان جمع‌آوری و به میزبان مرکزی ارسال می‌شود. در میزبان مرکزی عملیات پیچیده‌تر انجام می‌شود. سامانه‌های تشخیص رخنه مبتنی بر میزبان از رخدادهای سامانهٔ و اطلاعات مربوط به برنامه‌های کاربردی به‌عنوان منابع اطلاعاتی استفاده می‌کنند. منابع اطلاعاتی که در سامانه‌های مبتنی بر میزبان نظارت می‌کنند، عبارت‌اند از [۳۲]:

[1]Reactive IDS
[2]Internet Protocol

[3]Signature-baseddetection
[4]Completly
[5]Host Intrusion Detection System

- سامانه پرونده
- رویدادهای شبکه
- فراخوانی‌های شبکه

سامانه‌های مبتنی بر میزبان فعالیت‌های مجاز بر روی رایانه میزبان را شناسایی و حملات را تشخیص می‌دهند.

- **سامانه‌های مبتنی بر شبکه**[1]

در این سامانه‌ها، بر پرونده‌ها و داده‌های مربوط به ترافیک شبکه، سایر اجزای شبکه و سامانه‌های کنترل امنیت مانند دیوارهای‌آتش نظارت دارند. سامانه‌های مبتنی بر شبکه، بسته‌های عبوری در سطح شبکه را به‌عنوان منبع اطلاعات جمع‌آوری می‌کنند و احتمال آسیب‌رساندن به شبکه را کاهش می‌دهند. سامانه‌های تشخیص رخنه مبتنی بر شبکه، اطلاعات ترافیک شبکه را مورد تجزیه‌وتحلیل قرار می‌دهند [۳۲،۲۴].

پروژه DIDS[2]، توسط انجمن‌ها و سازمان‌های امنیتی و نظامی آمریکا تحت حمایت قرار گرفت و زیر نظر مستقیم نیروی هوایی، آژانس امنیت ملی و وزارت انرژی به انجام رسید. معماری این سامانه توسط استیو ماها[3] ارائه شد. نخستین تلاش در جهت استفاده از منابع و اطلاعات میزبان و شبکه به‌طور هم‌زمان، برای تشخیص رخنه بود[۲۵].

۳- تکنیک‌های یادگیری ماشین

تکنیک‌های یادگیری ماشین به توانایی یک برنامهٔ رایانه‌ای برای یادگیری و افزایش عملکرد آن در قالب مجموعه‌ای از وظایف در زمان را گویند. یادگیری ماشین به دو دسته، یادگیری با نظارت و یادگیری بدون نظارت تقسیم می‌شوند.

الف) یادگیری با نظارت: این الگوریتم به‌ازای هر ورودی دارای یک مقدار خروجی و یا تابع مشخص است. هدف سامانه یادگیری، به‌دست‌آوردن فرضیه‌ای است که تابع یا رابطهٔ بین ورودی و خروجی را حدس بزند. الگوریتم‌های یادگیری بانظارت به مجموعه‌ای از نمونه‌ها به نام مجموعهٔ آموزش نیاز دارند. شبکه‌های عصبی مصنوعی، الگوریتم‌های خوشه‌بندی k-means، شبکه‌های بیزین[4] و ماشین بردار پشتیبان ازجمله الگوریتم‌های یادگیری بانظارت می‌باشند [٦].

ب) یادگیری بدون نظارت: در این الگوریتم فقط مقدار ورودی‌ها مشخص است و اطلاعاتی در مورد خروجی در دسترس نیست. مقدار بعدی بر اساس موقعیت فعلی ورودی پیش‌بینی می‌شود[۲۰].

۳-۱- الگوریتم ماشین بردار پشتیبان

ماشین بردار پشتیبان یکی از روش‌های یادگیری بانظارت است. مبنای کار ماشین بردار پشتیبان بر مبنای دسته‌بندی خطی داده‌ها است. در تقسیم خطی داده‌ها سعی بر انتخاب خطی است که حاشیهٔ اطمینان بیش‌تری داشته باشد. ویژگی مهم ماشین بردار پشتیبان این است که برخلاف سایر شبکه‌های عصبی دیگر مانند MLP[5]، RBF به‌جای این‌که خطای مدل‌سازی و طبقه‌بندی را کم کنند، آن‌ها خطرپذیری عملیاتی را به‌عنوان تابع هدف در نظر می‌گیرند و مقدار بهینه آن را محاسبه می‌کنند. ماشین بردار پشتیبان با استفاده از کرنل غیرخطی قادر به تصمیم‌گیری غیرخطی نیز است. انتخاب کرنل مناسب برای ماشین بردار پشتیبان، منجر به برتری آن نسبت به سایر رویکردهای مبتنی بر تصمیم‌گیری خطی شده است [۱۰، ۸،۹] ؛ لذا استفاده از تکنیک‌های داده‌کاوی برای سامانه‌های تشخیص رخنه به این دلیل است که قادر به پیش‌بینی فعالیت‌های نفوذگران هستند [۷].

۳-۱-۱- انواع مدل‌های بردار پشتیبان

مدل‌های بردار پشتیبان به دو گروه عمده تقسیم می‌شوند:

- مدل طبقه‌بندی ماشین بردار پشتیبان[6]
- مدل رگرسیون بردار پشتیبان

مدل طبقه‌بندی ماشین بردار پشتیبان جهت حل مسائل طبقه‌بندی داده‌هایی که در طبقه‌های مختلف قرار می‌گیرند، استفاده می‌شود. هم‌چنین مدل رگرسیون پشتیبان نیز در حل مسائل پیش‌بینی کاربرد دارد. همان‌گونه که بیان شد، ماشین بردار پشتیبان مبتنی بر ساختار خطر است و از تئوری آموزش آماری برگرفته شده است [۱۱].

الگوریتم ماشین بردار پشتیبان در الگوریتم‌های تشخیص الگو دسته‌بندی می‌شود و در هرجایی که نیاز به تشخیص الگو یا دسته‌بندی اشیا در طبقه‌های خاص باشد، از آن استفاده می‌شود. در الگوریتم ماشین بردار پشتیبان، هدف، انتخاب بهترین خط است، ازاین‌جهت که کمترین میزان خطا برای طبقه‌بندی داده‌ها را داشته باشد. این روش را به الگویی

[1] Network intrusion detection system
[2] Distributed IDS
[2] Steve Maha
[4] Bayesian

[5] Multi Layer Perceptron
[6] Support vector classification

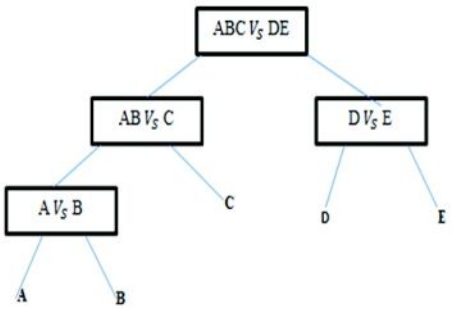
از فضای F که به‌طورمعمول دارای ابعاد بزرگی است، تبدیل می‌کند. هم‌چنین تعمیم خطای ماشین بردار پشتیبان به مشخصات هندسی داده‌های آموزشی نه به ابعاد ورودی بستگی دارد. ازجمله کاربردهای ماشین بردار پشتیبان، شبیه‌سازی مسیر، پیش‌بینی کیفیت، بهینه‌سازی زمان پیوند اعضا، تشخیص چهره، طبقه‌بندی ژن‌های سرطان و تجزیه‌وتحلیل بحران اقتصادی ارز و کاهش هزینه است [۳، ۱۳، ۱۲].

امروزه ماشین بردار پشتیبان به یک روش محبوب در مدل‌سازی تبدیل شده است که باعث کاهش زمان آموزش می‌شود. هدف استفاده از ماشین بردار پشتیبان در سامانه‌های تشخیص رخنه، جداکردن الگوهای طبیعی از نفوذی و هم‌چنین دارای عملکرد بهتری در تشخیص حملات DOS و حملات کاوشی[1] (Prob) است [۱۵، ۱٤].

۳-۲- به‌کارگیری ماشین‌های بردار پشتیبان با دیگر تکنیک‌های داده‌کاوی

درخت‌های تصمیم روشی برای نمایش یک سری از قوانین هستند که منتهی به یک رده یا مقدار می‌شوند. یکی از تفاوت‌های بین روش‌های ساخت درخت تصمیم این است که این فاصله چگونه اندازه‌گیری می‌شود. درخت‌های تصمیمی که برای پیش‌بینی متغیرهای دسته‌ای استفاده می‌شوند، درخت‌های طبقه‌بندی[2] نامیده می‌شوند؛ زیرا نمونه‌ها را در دسته‌ها یا رده‌ها قرار می‌دهند. درخت‌های تصمیمی که برای پیش‌بینی متغیرهای پیوسته استفاده می‌شوند، درخت‌های رگرسیون نامیده می‌شوند [۲۶].

ماشین‌های بردار پشتیبان بر پایه درختان تصمیم‌گیری یکی از روش‌هایی است که مولای[3] و همکارانش به آن اشاره کردند [۳۳]. بدین ترتیب در جهت حل مسائل از داده‌ها با طبقه‌های گوناگون استفاده می‌شود. ترکیب مدل‌های مختلف، سرعت و عملکرد بهتری در مقایسه با مدل‌های تصمیم‌گیری با مدل‌های یادگیری به‌صورت منفرد دارد؛ در این مقایسه از پنج مدل ماشین بردار پشتیبان برای پنج گروه از داده‌های برچسب‌گذاری‌شده استفاده می‌شود. این داده‌ها شامل چهار گروه از حملات و یک گروه از داده‌های سالم یا عادی استسپس ، پنج الگوی استخراج‌شده در هر داده را به دو طبقه عادی و حمله تقسیم می‌کند. بنابراین در مراحل بعدی نیز همین چرخه برای چهار نوع حملهٔ دیگر انجام می‌شود. با توجه به درخت زیر ، پنج گروه از داده‌ها به‌صورت A.B.C.D.E فرض شده است. در شکل (۳-۱۶) ابتدا دو گره از داده‌ها که بیش‌ترین فاصله را از هم دارند به دو گروه جدا تقسیم می‌شوند؛ به‌عنوان‌مثال A تا E به دو گروه تقسیم می‌شوند. فاصلهٔ سایر گروه‌ها یعنی B و C و D ازA و E اندازه گرفته می‌شود و هرکدام از این دو که نزدیک‌تر بود در گروه آن‌ها قرار می‌گیرد. همان‌طور که در ریشهٔ درخت مشاهده می‌کنید، فاصلهٔ B و C به A نزدیک‌تر بوده است. بنابراین در یک گروه قرار می‌گیرند و فاصلهٔ D به E نزدیک‌تر بوده است. بنابراین این دو نیز در یک گروه قرار می‌گیرند. این فرآیند برای تمام زیردرخت‌ها ادامه می‌یابد تا تمام گره‌ها از هم جدا شوند، بدین‌ترتیب پنجم گروه از داده‌ها از هم جدا می‌شوند و امکان طبقه‌بندی و تشخیص وجود دارد. بااین‌وجود از ماشین‌های بردار پشتیبان دودویی به‌کمک درخت‌های تصمیم‌گیری برای حل مسئله‌ای با چند طبقه از داده‌ها استفاده می‌شود. البته هدف، کاهش زمان آموزش مدل‌ها و افزایش عملکرد تشخیص است.



(شکل ۱): روش جداسازی داده‌ها[۳۳].

ژانگ[4] و وانگ برای بهینه‌سازی ساختار کرنلی ماشین بردار پشتیبان، ویژگی‌های اضافی و اثرات منفی را کشف و شناسایی کردند. همچنین برای بهینه‌سازی اطلاعات کرنل ماشین بردار پشتیبان، کارایی پارامتر تابع یادگیری ماشین بردار پشتیبان را افزایش دادند که این امر منجر به حل مشکلات در محاسبات می‌شود[۲۷].

استفاده از الگوریتم ژنتیک و ماشین بردار پشتیبان نیز برای تشخیص خودکار مجموعه‌ای از ویژگی‌ها مناسب است. انتخاب ویژگی‌های مهم به‌منظور افزایش قابلیت تعمیم و

[1] Probing
[2] Classification
[3] Mulay
[4] Xiong

سرعت‌بخشیدن به فرآیند یادگیری و بهبود الگو است. محاسبهٔ فرآیند فیتنس[1]، انتخاب کراس آور[2] و جهش برای اجرای حداکثر تعداد نسل میسر است. با نخبه‌گرایی در هر نسل، بهترین رشته برای آن نسل ایجاد می‌شود. بنابراین محل به‌دست‌آمده، بهترین مجموعه دسته‌بندی است[۲۸]. مدل پیشرفته ماشین بردار پشتیبان با کرنل تابع وزن‌دار، دارای عملکرد بهتری، نسبت به ماشین بردار پشتیبان معمولی است. الگوریتم جدید ماشین بردار پشتیبان با توجه به سطوح وزنی برای رتبه‌بندی ویژگی‌های تشخیص رخنه داده مورد استفاده قرار می‌گیرند. برای تشخیص رخنه، دقت، زمان محاسبه و میزان خطا بسیار کم‌تر است. شکل کلی کرنل تابع جدید به‌صورت رابطهٔ (۱)

$$f(x) = \mathrm{sign}\left(\sum_{i=1}^{L} \alpha_i y_i k(w_{xi}, w_x) + b\right) \qquad (۱)$$

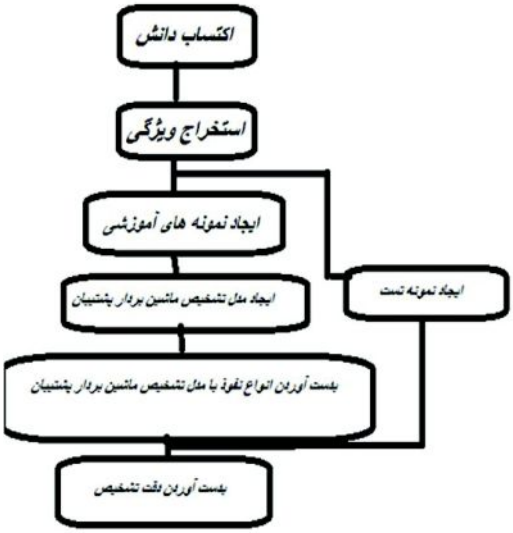
تابع ذکرشده برای پیداکردن وزن با حداکثر حاشیهٔ ابر صفحه به‌منظور جداکردن دو طبقه بر مبنای تفکیک‌کننده غیرخطی است. در رابطهٔ (۱)، L تعداد رکوردهای آموزش و $y_i \in \{-1,1\}$ برچسب اطلاعات آموزش داده شده است که $0 \leq \alpha_i \leq C$ و بردار پشتیبان بر اساس وزن است[۲۹].

پیاده‌سازی سامانه‌های تشخیص رخنه با استفاده از ماشین بردار پشتیبان احتمال خطر، خطای تعمیم و خطا در آموزش را کاهش می‌دهد. روش PCNN[3] (اجزای اصلی شبکه‌های عصبی) مبتنی بر الگوریتم تشخیص رخنه، هشدارهای اشتباه را کاهش می‌دهد. هم‌چنین استخراج ویژگی از طریق PCNN، عملکردهای تشخیص رخنه مبتنی بر ماشین بردار پشتیبان را نیز بهبود می‌بخشد. بااین‌وجود استخراج ویژگی، کارآمدترین روش برای تقویت عملکرد سامانه تشخیص رخنه است[۳۰].

۳-۳- فرآیند تشخیص رخنه با استفاده از ماشین بردار پشتیبان

روند تشخیص رخنه توسط ماشین بردار پشتیبان در شکل ۲ نشان داده‌شده است. با استخراج ویژگی، داده‌ها به نمونه‌های آموزشی و نمونه‌های آزمون تقسیم می‌شوند. با توجه به این‌که نمونه‌های آموزشی برای تشخیص، توسط ماشین بردار پشتیبان و نمونه‌های آزمون برای به‌دست‌آوردن دقت تشخیص از مدل ماشین بردار پشتیبان استفاده می‌کنند[۳۱].

[1]Fitness
[2]Crossover
[3]Principal Component Neural Networks



(شکل ۲): فرآیند تشخیص رخنه توسط ماشین بردار پشتیبان[۳۱]

۳-۴- روند کار ماشین بردار پشتیبان

حل معادله پیداکردن خط بهینه برای داده‌ها، به‌وسیله روش‌های برنامه‌نویسی درجهٔ دوم[4] ازجمله روش‌های شناخته‌شده‌ای است که به‌صورت محدود شده می‌باشد.

۳-۴-۱- مدل طبقه‌بندی ماشین بردار پشتیبان

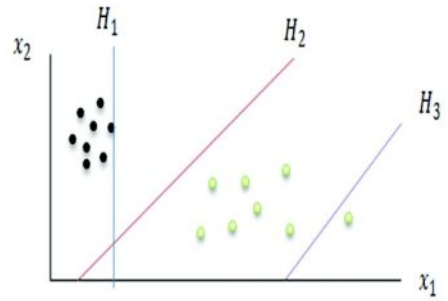
فرض کنید تعدادی از بردارهای ویژگی یا الگوهای آموزشی به‌صورت $\{x_1, x_2 \ldots, x_i, \ldots, x_N\}$ وجود داشته باشد که هرکدام یک بردار ویژگی d بعدی می‌باشند. با برچسب y_i به‌طوری‌که $y_i \in \{-1, +1\}$ است.

(شکل ۳): تعدادی نمونه آموزشی در فضای دوبعدی.

پایه معادلات ماشین بردار پشتیبان به‌صورت دوطبقه تعریف می‌شود که هدف،حل یک مسئله دسته‌بندی دوطبقه به‌صورت بهینه است. مجموعه داده‌ها به‌وسیله خط از یکدیگر جدا می‌شوند. ماشین بردار پشتیبان، خطرپذیری عدم طبقه‌بندی را به‌صورت یک‌سری کمیت‌های عددی بیان و سپس مقدار

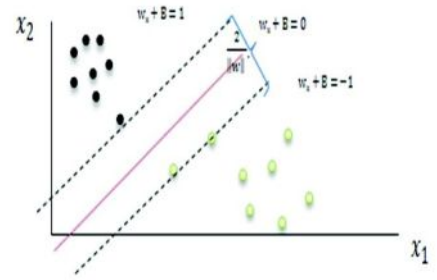
[4] Quadratic Programming

کمینه را از بین آنها حساب می‌کند؛ اما شبکه عصبی MLP زمانی که تفکیک درست باشد، مقدار خطای تفکیک را نیز کمینه می‌کند و آن را به‌عنوان یک راه حل در نظر می‌گیرد. ازآنجاکه مجموعه‌داده‌های آموزشی دارای برچسب می‌باشند و به‌وسیلهٔ تفکیک‌کننده، نقاط متعلق به دو طبقه از یکدیگر تفکیک می‌شوند؛ به این تفکیک‌کننده مرز تصمیم‌گیری گفته می‌شود. شکل۴ به تعدادی از مرزهای تصمیم‌گیری اشاره دارد.



(شکل ۴): مرز تصمیم‌گیری[۱۶] .

در شکل ۴ تفکیک‌کننده H_3، دو طبقه را به‌درستی طبقه‌بندی نمی‌کند؛ بنابراین تفکیک‌کننده H_3، تفکیک‌کننده مطلوبی نیست؛ اما دو تفکیک‌کننده H_1 و H_2 دارای خطرپذیری عملیاتی هستند که H_2 نیز دارای خطرپذیری عملیاتی کمتری نسبت به H_1 است. به عبارتی H_2 در شکل بالا، نقطهٔ تعادل سامانه است. به هر یک از اعضای طبقه، بردار پشتیبان گویند. نزدیک‌ترین عضو در هر طبقه سعی بر دورکردن خط تفکیک‌کننده از خود را دارد. حاشیه مرز به کوتاه‌ترین فاصله میان نزدیک‌ترین نقطه در هر دو طرف نیم‌صفحه اشاره می‌کند. در شکل ۵ به‌طور بدیهی قابل شهود و از نظر ریاضی نیز قابل اثبات است که تفکیک‌کننده، حداکثر خطا را پیدا می‌کند [۱۶].



(شکل ۵): تفکیک‌کننده مطلوب [۱۶] .

۳-۵- مدل رگرسیون بردار پشتیبان خطی

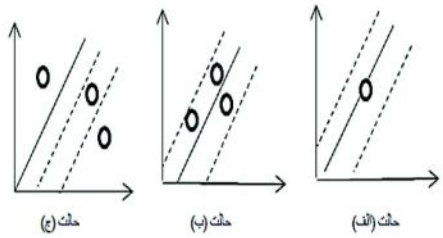
یکی دیگر از انواع ماشین‌های بردار پشتیبان، رگرسیون بردار پشتیبان است. در سال ۱۹۹۷، وپنیک با همکاری Steven Golowich و Alex Smola، رگرسیون بردار پشتیبان را پیشنهاد دادند[۳۵]. رگرسیون بردار پشتیبان، درواقع تعمیمی است بر آن ایده‌ای که ماشین‌های بردار پشتیبان برای طبقه‌بندی دودویی دارند. ماشین‌های بردار پشتیبان مبتنی بر ورودی‌ها در فضای d بعدی هستند و خروجی‌های آن در حالت کلی دو حالت است؛ اما در اینجا فرض بر این است که خروجی‌های ما بیش از دو مقدار و در حالت کلی بی‌نهایت مقدار است. به‌عبارتی خروجی‌ها به شکل حقیقی هستند. رگرسیون بردار پشتیبان در زمینه‌های مختلف سری‌های زمانی و مالی، پیش‌بینی، تقریب و تجزیه‌وتحلیل‌های پیچیده و انتخاب توابع به‌کارگرفته می‌شود.

رابطهٔ (۲) تفاوت ماشین بردار پشتیبان و رگرسیون بردار پشتیبان را نشان می‌دهد. در این رابطه، x_i نمایان‌گر ورودی در فضای d بعدی در ماشین بردار پشتیبان و رگرسیون بردار پشتیبان است. t_i نمایان‌گر خروجی است که در ماشین بردار پشتیبان، خروجی‌ها به‌صورت دودیی (۱+ و ۱-) نشان داده می‌شوند؛ اما در رگرسیون بردار پشتیبان، خروجی‌ها در بازهٔ مجموعه اعداد حقیقی هستند.

رابطه (۲)

$$\mathrm{SVM}(x_i, t_i)\begin{cases} x_i \in \mathbb{R}^d \\ t_i \in \{-1, +1\} \end{cases} \quad \mathrm{SVR}(x_i, t_i)\begin{cases} x_i \in \mathbb{R}^d \\ t_i \in \mathbb{R} \end{cases}$$

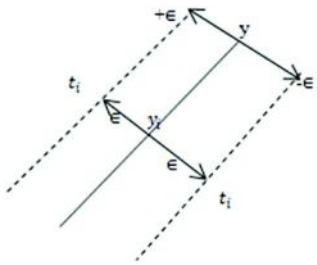
درصورتی‌که یک نقطه وجود داشته باشد، معادلهٔ خط دقیقی که از آن می‌گذرد $y = w^T x + b$ است و دو خط تقریبی در آن وجود دارد. هدف در اینجا استفاده از این خط‌های تقریبی است. حال با فرض این‌که به‌جای یک نقطه، چند نقطه داشته باشیم، در رگرسیون بردار پشتیبان باید نواری را به دست آوریم که این نقاط را پوشش دهند و نقاط در آن قرار گیرند. در شکل ۶ بین دو شکل (ب) و (ج)، شکل (ب) بهتر است که با یک نوار خطی می‌توان سه نقطه را پوشش داد؛ اما در شکل (ج)، نوار تفکیک‌کنندهٔ مناسبی نیست؛ بنابراین هدف، پیداکردن بهترین تفکیک‌کننده‌ای است که بهترین عمل پوشش را انجام دهد.



(شکل ۶): یافتن بهترین خط تقریبی در رگرسیون بردار پشتیبان .

بنابراین در شکل (ج) نقاطی که در نوار قرار ندارند، باید جریمه شوند.

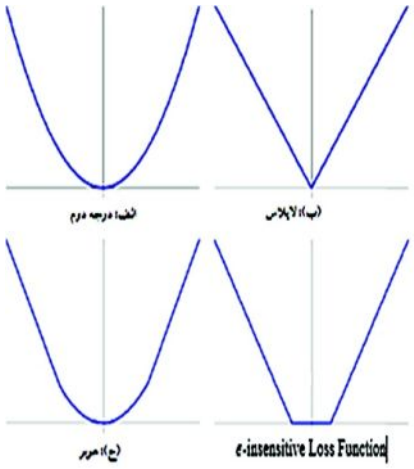
تفاوت t_i و y_i در شکل (۷) این است که t_i مقدار تقریبی، اما y_i یک مقدار دقیق است. یعنی با فرض این‌که رگرسیون خطی $t_i \approx y_i = w^T x_i + b$ باشد.



(شکل ۷): بازهٔ مطلوب تفکیک‌کننده در رگرسیون بردار پشتیبان.

به این نکته باید توجه کرد که هرچه $\|w\|$ کم‌تر باشد مدل، یک مدل ساده‌تری است. با وجود بی‌شمار نقطه، باید تفکیک‌کننده‌ای را پیدا کرد که دارای تقریب دقیق‌تری باشد. با توجه به شکل ۷ درصورتی یک نقطه در نظر گرفته شود و خطی که از آن می‌گذرد، رسم شود. اگر نقطه به‌طوردقیق روی خط نباشد، مشکلی وجود ندارد و می‌توان این خطا را تحمل کرد؛ به این حد ϵ می‌گویند. این بازه به‌عنوان یک بازهٔ مطلوب است؛ یعنی مادامی‌که داده‌ها در این نوار هستند، این مدل برای داده‌ها مورد قبول است؛ اما اگر داده‌ها خارج از این فضا باشند، خطای آن بیش‌تر از ϵ می‌شود. درنتیجه باید آن‌ها را به‌نحوی جریمه کرد. چهار نوع از توابع جریمه در شکل ۶ نشان‌داده شده است.

تابع جریمه در شکل ۸ قسمت (الف) مربوط به معیار خطا در حداقل مربعات است. تابع جریمه در قسمت (ب) تابع جریمهٔ لاپلاس[1] است که به نقاط دورافتاده حساس است. تابع جریمه در قسمت (ج) به‌عنوان تابع جریمهٔ هوبر به‌عنوان تابع جریمهٔ قوی محسوب می‌شود و تابع جریمه در قسمت (د) تابع جریمه[2] L_ϵ است که دارای خواص زمانی مطلوب است [۱۷]. درواقع این میزان جریمه یک تابع جریمه به شکل L_ϵ است.



(شکل ۸): انواع توابع جریمه [۱۷].

۳-۶- تئوری مجموعه‌های راف

تئوری مجموعه‌های راف یا آر. اس. تی[3] توسط پژوهش‌گر لهستانی به نام پاولاک در سال ۱۹۸۲ معرفی شد [۱۸]. ازآنجاکه پژوهش‌های نخستین در زمینهٔ تئوری مجموعه‌های راف به زبان لهستانی منتشر شد؛ لذا نظر پژوهش‌گران علوم رایانه و ریاضیات در سطح بین‌الملل را جلب نکرد و توجه چندانی به این تئوری صورت نگرفت. در اواخر دههٔ هشتاد این تئوری در سطح جهانی مطرح شد و در سال ۱۹۹۲ نخستین کنفرانس علمی تئوری مجموعه‌های راف به‌صورت بین‌المللی در لهستان برگزار شد . در سال ۱۹۹۵ از سوی ACM[4] به‌عنوان موضوع نوظهور در علم رایانه معرفی شد. ازآن‌پس تئوری مجموعه‌های راف به‌عنوان یک ابزار محاسباتی جدید برای برخورد با شرایط مبهم و عدم قطعیت شناخته شد. این امر می‌تواند برای تحلیل اطلاعات غیردقیق متناقض و ناکامل به کار گرفته شود. امروزه تئوری مجموعه‌های راف به‌عنوان ابزاری قدرتمند برای استنتاج داده‌ها و تحلیل و پیش‌بینی تصمیمات داده‌کاوی سامانه‌های خبره و سامانه‌های پشتیبانی تصمیم و بسیاری از زمینه‌های دیگر شناخته شده است.

[1] Laplace

[2] ϵ-insensitive Loss Function

[3] Rough Set Theory

[4] Association for computing machinery

تئوری مجموعه‌های راف را می‌توان به تحلیل‌های تصمیم‌گیری و تحلیل‌های غیر تصمیم‌گیری تقسیم کرد. تحلیل‌های غیر تصمیم‌گیری به‌طور عمده، فشرده‌کردن اطلاعات، تقلیل اطلاعات، خوشه‌بندی، کشف الگو و نظایر آن را شامل می‌شود. درواقع کارکرد اصلی این دسته از تحلیل‌ها آن است که ویژگی‌های غیرضروری را حذف و با فشرده‌کردن و تقلیل داده‌ها، امکان تحلیل بهتر داده‌ها را فراهم کنند. تحلیل‌های تصمیم‌گیری نیز به کشف و استخراج قوانین تصمیم کمک می‌کنند [۳۶]. این تئوری با تحلیل جدول‌های داده سروکار دارد. مجموعه راف یکی از ابزارهای ریاضی است که قادر به توجیه عدم قطعیت و ابهام است. مجموعه راف الگوهای خاص در داده‌های ناقص را از منابع اطلاعاتی کشف می‌کند؛ همچنین برای طیف گسترده‌ای از برنامه‌های کاربردی مربوط به تشخیص الگو، پردازش تصویر، انتخاب ویژگی، محاسبات عصبی، به‌عنوان تجزیه‌وتحلیل کنندهٔ چند منبع داده، کشف دانش از داده‌هایی با ابعاد بالا و سامانه‌های اطلاعاتی توزیع و پشتیبانی تصمیم‌گیری استفاده می‌شود [۱۸].

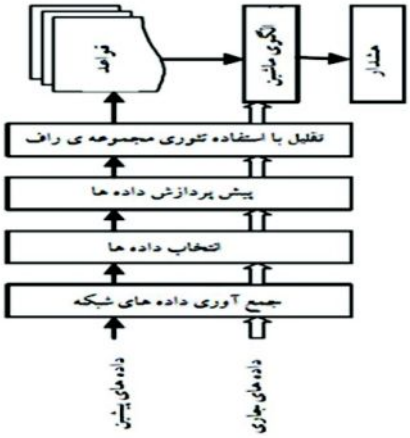
تشخیص رخنه مبتنی بر تئوری مجموعهٔ راف فرایندی تعاملی و تکرارشونده شامل مراحل زیر است:

۱- جمع‌آوری داده‌های شبکه

۲- انتخاب داده‌ها: هدف از این مرحله انتخاب جداول، سامانه اطلاعاتی و ویژگی‌های حملات است.

۳- پیش‌پردازش داده‌ها: این مرحله شامل پردازش اطلاعات ناقص است. به‌عنوان‌مثال، برخی داده‌ها که اطلاعات کافی را ندارند، با استفاده از تئوری مجموعهٔ راف سوابق ناقص آن حذف می‌شوند.

۴- تقلیل: این مرحله در تئوری مجموعهٔ راف منجر به تولید قوانین و الگوهای طبیعی و غیرطبیعی در شبکه می‌شود. اگر مشخصهٔ یک زیرمجموعه $B \subset A$ باشد از ویژگی‌های A-B چشم‌پوشی می‌شود. تقلیل، مانند زیرمجموعه‌ای است که شامل حداقل ویژگی‌های غیرضروری نباشد. بااین‌وجود، سامانه‌های تشخیص رخنه با استفاده از تئوری مجموعهٔ راف منجر به تشخیص دقت، پردازش داده و کاهش هشدار اشتباه می‌شود [۹].



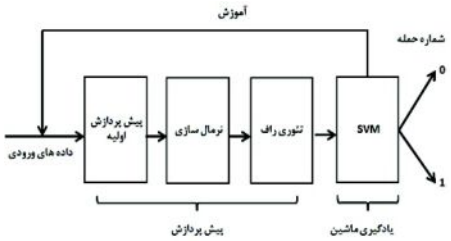
(شکل ۹): سامانه تشخیص رخنه مبتنی بر تئوری مجموعهٔ راف [۱۹].

۴- یافته‌های پژوهش و الگوریتم پیشنهادی

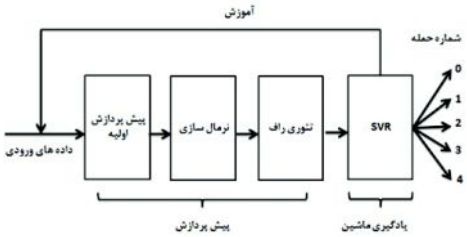
الگوریتم پیشنهادی برای سامانه تشخیص رخنه در شبکه‌های رایانه‌ای مورد استفاده قرار می‌گیرد. در این الگوریتم با استفاده از یادگیری مبتنی بر ماشین و پیش‌پردازش بر اساس تئوری راف بررسی می‌شود که یک سامانه ترکیبی بر اساس درجهٔ همبستگی برحسب مهم‌ترین ویژگی‌ها در حملات به‌کار گرفته می‌شود تا دقت و سرعت یادگیری ماشین بیش‌تر شود. در روش‌های مشابه با یادگیری ماشین روی تمام ویژگی‌ها، عملیات پردازش انجام می‌گیرد. با استفاده از این روش بر اساس تمام ویژگی‌های ترافیک شبکه عصبی مبتنی بر یادگیری ماشین، تصمیم می‌گیرد که کدام دادهٔ آزمون متعلق به کدام خوشه (حمله) است. به‌کارگیری این روش‌ها خالی از ایراد نیست. یکی از مهم‌ترین ایرادها، استفاده از ویژگی‌هایی است که در تعیین نوع حمله نقشی ندارند. این امر منجر به کاهش دقت یادگیری ماشین می‌شود. از طرف دیگر پردازش، به‌کارگیری داده‌ها و ویژگی‌هایی که نقش مهمی در طبقه‌بندی ندارند، تشخیص رخنه و سرعت یادگیری را نیز کاهش می‌دهند. در این پژوهش سعی کردیم با یک روش جدید، تئوری راف را با یادگیری ماشین ترکیب کنیم. ازآنجاکه این تئوری با تحلیل جدول‌های داده سروکار دارد و جدول‌های داده نیز توسط اندازه‌گیری یا افراد متخصص و آگاه حاصل می‌شوند، هدف اصلی از تحلیل مجموعه راف به دست‌آوردن مفاهیم تقریبی از داده‌های اکتسابی است. این

تئوری، یک ابزار قدرتمند ریاضی برای استدلال در موارد ابهام و غیرقطعی است که روش‌هایی را برای زدودن و کاستن اطلاعات دانش نامربوط یا مازاد بر نیاز از پایگاه‌های داده‌ها مهیا می‌سازد. در این فرایند حذف داده‌های زائد بر مبنای آموزش، وظیفهٔ اصلی سامانه است و بدون از دست‌دادن داده‌های اساسی پایگاه داده‌ها صورت می‌پذیرد. بنابراین از تقلیل اطلاعات، مجموعه‌ای از قواعد تلخیص‌شده و پرمعنا حاصل می‌شود که کار تصمیم‌گیرنده را بسیار ساده‌تر می‌کند. درحقیقت می‌توان گفت که مجموعه راف با کاهش فضای داده‌ها و برگزیدن ویژگی‌های مهم، یک نگاشت از فضای داده‌های خام را به فضای سمانتیک (مفاهیم) انجام می‌دهد. الگوریتم پیشنهادی با استفاده از تئوری راف تمام ویژگی‌های یک حمله را به تعدادی از مهم‌ترین ویژگی‌های حملات نگاشت می‌دهد.

درواقع در مرحله پیش‌پردازش علاوه‌بر هنجارسازی، درجه‌ای از میزان همبستگی را بر اساس تئوری راف به کار می‌بریم که بر اساس مهم‌ترین ویژگی‌ها در یادگیری ماشین، نظیر ماشین بردار پشتیبان به کار گرفته می‌شود. در شکل ۱۰ دیاگرام کلی سامانهٔ پیشنهادی برای دسته‌بندی در دو طبقه (حمله) معرفی می‌شود. البته به‌دلیل این‌که اگر تعداد طبقه‌های حملات بیشتر از دو عدد باشد (در روش پیشنهادی پنج دسته داریم) بهتر است از حالت چند طبقهٔ ماشین بردار پشتیبان (رگرسیون بردار پشتیبان) استفاده کنیم. در شکل ۱۱ این سامانهٔ پیشنهادی با پنج طبقه‌بندی حمله نشان داده شده است.

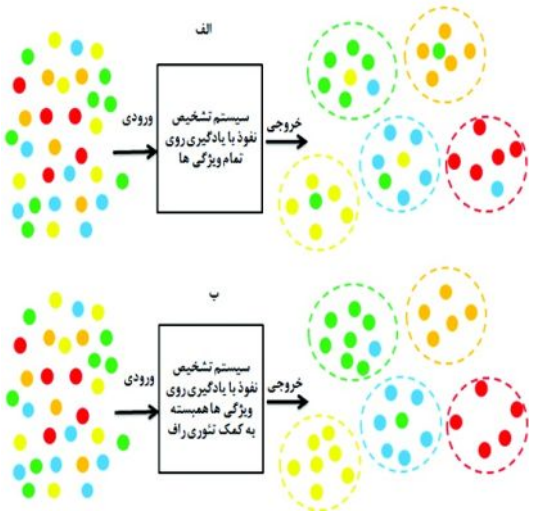


(شکل ۱۰): سامانه پیشنهادی تشخیص رخنه با دو طبقه‌بندی.



(شکل ۱۱): سامانه پیشنهادی تشخیص رخنه با پنج طبقه‌بندی.

در این پژوهش ، از دادگان ۱۰ درصد از مجموعه پوشه‌های دادگان KDD CUP 99 استفاده شده است. برای آزمایش از پنجاه درصد کل دادگان، حدود هفتاد درصد به داده‌های آموزشی و سی درصد به داده‌های آزمون اختصاص داده شده است. مجموعه یادگیری شامل دو مجموعه است: مجموعه نخست از تمام ویژگی‌ها و مجموعه دوم یادگیری، فقط برحسب ویژگی‌هایی است که در تئوری راف عنوان شده است. برای مجموعه آموزشی نخست زمان زیادی صرف یادگیری می‌شود و یادگیری ماشین مجبور به استفاده از ویژگی‌هایی است که همبستگی زیادی با حملات ندارند. این یادگیری درواقع یادگیری با داده‌های پرت است و تأثیر منفی روی دقت سامانهٔ تشخیص رخنه می‌گذارند. در مجموعهٔ آموزشی دوم فقط از ویژگی‌هایی که در تئوری راف ثابت شده است که دارای همبستگی بالایی با حملات هستند، استفاده می‌شود. استفاده از این روش، منجر به کاهش ابعاد مجموعهٔ آموزشی و افزایش سرعت تشخیص رخنه می‌شود. به‌علت حذف ویژگی‌های پرت (مانند حذف نوفر)، در اکثر موارد دقت تشخیص رخنه بالا می‌رود. در شکل (۱۲) نمایی از این روش نمایش داده می‌شود:



(شکل ۱۲): الف. خوشه‌بندی سامانه تشخیص رخنه با استفاده از یادگیری ماشین روی تمام ویژگی‌ها را نشان می‌دهد در قسمت ب. یادگیری ماشین بجای استفاده از تمام ویژگی‌ها از ویژگی‌های مهمی که در تئوری راف به‌دست‌آمده استفاده کرده است با این روش سرعت آموزش و تشخیص بالا می‌رود و به‌علت حذف ویژگی‌های کم‌اهمیت میزان خطا در خوشه‌بندی کاهش یافته است. (در شکل ب. خوشه‌بندی دقیق‌تری نسبت به حالت الف ایجادشده است).

۵- ارزیابی روش پیشنهادی

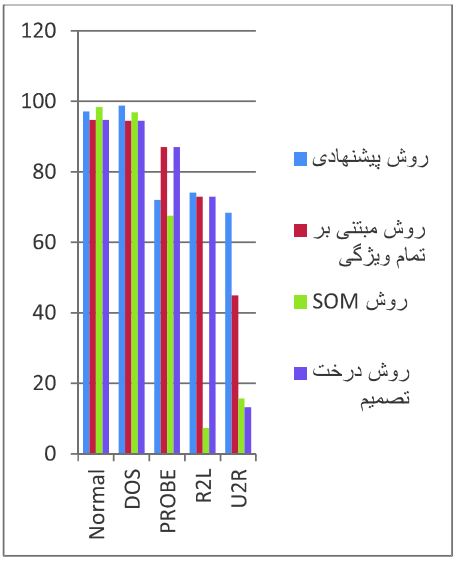
جدول ۱ میزان دقت تشخیص الگوریتم پیشنهادی با سه روش تشخیص رخنه برای پنج حمله اصلی را مقایسه می‌کند. در این مقایسه‌ها تعداد رکوردهای آموزشی دو هزار رکورد است. از اطلاعات جدول ۱ مشخص است که الگوریتم پیشنهادی در تشخیص حملات Normal بهتر از الگوریتم یادگیری بر مبنای تمام ویژگی‌ها و درخت تصمیم‌گیری است؛ ولی اندکی از تشخیص رخنهٔ شبکهٔ عصبی SOM کمتر است. بنابراین میزان دقت الگوریتم پیشنهادی در تشخیص حملات DOS از تمام الگوریتم‌های جدول ۱ بهتر عمل می‌کند؛ اما میزان دقت تشخیص حملات Probe از تشخیص رخنه شبکه عصبی SOM بیشتر است. در مورد حملات R2l و U2R نیز نتایج الگوریتم پیشنهادی از سه روش دیگر بهتر عمل می‌کنند.

(جدول ۱): محاسبه دقت تشخیص در چهار روش مورد بررسی در پژوهش.

نام حمله	درخت تصمیم‌گیری	شبکه عصبی SOM	تمام ویژگی‌ها	روش پیشنهادی
Normal	۹۴/۵۰	۹۸/۴۰	۹۴/۷۳	۹۷/۱۴
Dos	۹۷/۱۰	۹۶/۹۰	۹۴/۵۱	۹۸/۸۳
Probe	۸۳/۳۰	۶۷/۶۰	۸۷	۷۲
R2l	۸/۴۰	۷/۳۰	۷۳	۷۴/۱۵
U2R	۱۳/۲۰	۱۵/۷۰	۴۵	۹۸/۴۲

نمودار شکل (۱۱) و (۱۲)، نمودار مقایسه دقت الگوریتم پیشنهادی، الگوریتم یادگیری بر پایه تمام ویژگی‌ها، شبکه عصبی SOM و درخت تصمیم‌گیری به‌ترتیب توسط خطوط پیوسته و نمودارهای میله‌ای نشان داده شده‌اند.

(شکل ۱۳): مقایسه دقت انواع حملات با الگوریتم یادگیری بر پایه روش پیشنهادی، تمام ویژگی‌ها، شبکه عصبی SOM.



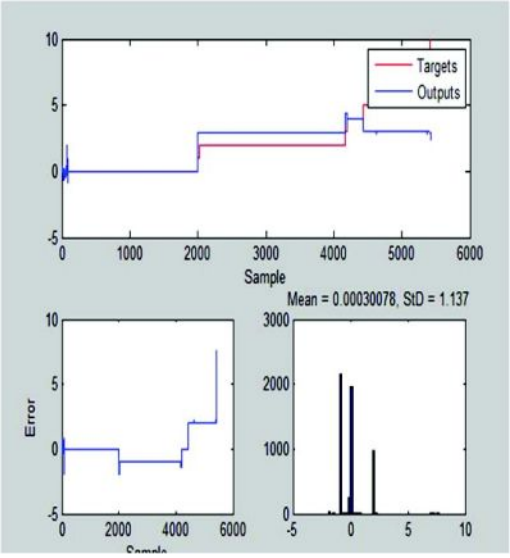
(شکل ۱۴): مقایسه دقت انواع حملات با الگوریتم یادگیری بر پایهٔ روش پیشنهادی، تمام ویژگی‌ها، شبکه عصبی SOM و درخت تصمیم گیری بر اساس نمودارهای میله‌ای.

از تحلیل نمودارهای شکل (۱۳) و شکل (۱۴) نشان داده می‌شود که دقت الگوریتم پیشنهادی از الگوریتم یادگیری بر پایه تمام ویژگی‌ها، شبکه عصبی SOM و درخت تصمیم‌گیری در حملات DOS، U2R و U2L بیشتر است. هم‌چنین تشخیص در حملات normal اندکی از روش SOM کمتر است؛ ولی در Probe، تشخیص رخنه حملات کمتر از روش‌های دیگر دارد. میزان دقت درصورتی بیش‌تر می‌شود که الگوریتم یادگیرنده بیشتر آموزش داده شود.

۵-۱- تشخیص حمله بر روی حملهٔ Back

الگوریتم روش پیشنهادی بر روی حمله Back پس از بیست‌بار آموزش به‌صورت شکل زیر است:

نمودار نخست در شکل ۱۵ ، حملهٔ نوع صفر را که به‌عنوان حمله Back شناخته می‌شود، نشان می‌دهد که مشتمل بر خطای زیادی نیست. خطای نوع صفر دارای دقت بالایی است. بنابراین حملات از صفر تا دو هزار از نوع حملهٔ صفر شناخته می‌شوند. با این وجود در نمودار دوم خطاهای نوع صفر به‌طور دقیق طبقه‌بندی می‌شوند. نمودار سوم نیز تعدادی از حملات را که درست تشخیص داده شده است در طبقه صفر تا دو هزار قرار داده است؛ همچنین میانگین خطا را نشان می‌دهد.



(شکل ۱۵): استفاده از روش پیشنهادی بر روی حمله Back.

۶- نتیجه‌گیری

در این پژوهش روشی مبتنی بر یادگیری ماشین، جهت طراحی یک سامانه تشخیص رخنه ارائه شده. یکی از ویژگی‌های شبکه‌های عصبی و سامانه‌های یادگیری ماشین، آموزش بر اساس داده‌های آموزشی است. این پژوهش از سامانه‌های تشخیص رخنه مبتنی بر یادگیری ماشین با خاصیت یادگیری روی ویژگی‌هایی که با تئوری راف دارای بیشترین ضریب همبستگی هستند، استفاده کرد. هم‌چنین از داده‌گان KDD CUP 99 استفاده می‌شود. این داده‌گان دارای تعداد زیادی رکورد که هر رکورد معرف یک ترافیک شبکه است . هر رکورد شامل ۴۱ ویژگی است. این داده‌گان از روی شبکه رایانه‌ای یکی از پایگاه‌های هوایی آمریکا که از نوع TCP است به مدت ۹ هفته گردآوری است. مدت‌زمان جمع‌آوری اطلاعات یادگیری حدود هفت هفته و مدت‌زمان گردآوری داده‌های آزمون حدود دو هفته به طول انجامید. این داده‌گان شامل ۲۲ حمله و یک حالت عادی است . این ۲۳ حمله، خود به پنج گروه اصلی DOS،U2R،R2L،Probe و normal تقسیم می‌شوند. سامانهٔ پیشنهادی متشکل از تعداد دو هزار رکورد به شکل تصادفی جهت آموزش شبکهٔ عصبی مبتنی بر یادگیری ماشین است. هم‌چنین تعدادی از رکوردهای داده‌گان به‌عنوان داده‌های آزمایشی استفاده شد. روش کار ما در این پژوهش به دو صورت می‌باشد: ابتدا از دو هزار رکوردی که برای آموزش در نظر گرفته شده اس،؛ تنها رکوردهایی را که با استفاده از تئوری راف ، بیش‌ترین همبستگی را به یک حملهٔ خاص دارند، برای آموزش در نظر گرفتیم. به‌عبارتی این روش پیشنهادی، یادگیری را روی تمام ویژگی‌ها انجام نمی‌دهد؛ بلکه بر روی ویژگی‌هایی که دارای ضریب همبستگی بیش‌تری هستند انجام می‌شود. این تکنیک، ابعاد ویژگی را کاهش می‌دهد؛ این کاهش باعث می‌شود سرعت یادگیری و تشخیص رخنه افزایش یابد. با این سازوکار، سامانهٔ تشخیص رخنهٔ پیشنهادی ما به سمت یک سامانهٔ تشخیص رخنه با سرعت بالا در زمان اجرا حرکت می‌کند. ما در این پژوهش میزان دقت روش پیشنهادی را با الگوریتم یادگیری بر پایه تمام ویژگی‌ها، شبکهٔ عصبی SOM و درخت تصمیم‌گیری مقایسه کردیم. نتایج نشان داد که میزان دقت الگوریتم پیشنهادی از الگوریتم یادگیری بر پایه تمام ویژگی‌ها، شبکهٔ عصبی SOM و درخت تصمیم‌گیری در حملات DOS،U2R و R2L بیشتر است و در حملات normal اندکی تشخیص آن از روش SOM کمتر است؛ ولی در حملات Probe میزان دقت، کمتر از روش‌های دیگر است. روش پیشنهادی به‌علت کاهش ابعاد یادگیری، دارای سرعت تشخیص رخنه بیش‌تری نسبت به الگوریتم یادگیری بر پایه تمام ویژگی‌ها است. الگوریتم پیشنهادی، مقدار DR بیشتری نسبت به روش الگوریتم یادگیری بر پایه تمام ویژگی‌ها دارد و مقدار FAR روش پیشنهادی اندکی بیشتر از الگوریتم یادگیری بر پایه تمام ویژگی‌ها است.

۷- پژوهش‌ها و پیشنهادهای آینده

در این پژوهش نشان دادیم که الگوریتم پیشنهادی دارای سرعت بالایی برای آموزش و تشخیص رخنه است. دقت الگوریتم پیشنهادی ما بالا است؛ ولی برای کاربردهای عملی پیشنهاد می‌کنیم که الگوریتم پیشنهادی ما به‌عنوان ورودی شبکه قرار گیرد و درصورتی‌که ترافیکی را به‌عنوان حمله تشخیص داد؛ این ترافیک را هشدار اعلام نکند؛ سپس ترافیک را تحویل یک سامانهٔ تشخیص رخنهٔ مبتنی بر یادگیری ماشین با داده‌های آموزشی زیاد کند. با این سازوکار اگر ترافیکی ازنظر سامانهٔ پیشنهادی، حمله تشخیص داده شود، هشدارها فعال نمی‌شوند؛ بلکه یک شانس دیگر به این ترافیک داده می‌شود که در یک سامانه سخت‌گیرانه‌تر عادی‌بودن خود را ثابت نماید. برتری این سامانهٔ ترکیبی این است که سرعت کلی سامانه تشخیص رخنه برای ترافیک‌های عادی کم نمی‌شود، از طرفی در این سامانه پیشنهادی ترکیبی، تعداد هشدارهای اشتباه کم می‌شود، بدون اینکه سرعت تشخیص کاهش یابد. درواقع از

یک فیلتر دولایه برای حملات استفاده می‌کنیم. لایهٔ نخستین این فیلتر، سامانه پیشنهادی این پژوهش است. اگر در لایهٔ نخستین، ترافیک به‌عنوان حمله تشخیص داده شد این ترافیک مسدود و تحویل لایهٔ دوم داده می‌شود. اگر لایهٔ دوم احساس کرد که این حمله نیست، از حالت مسدود خارج و در غیر این صورت هشدارهای لازم صادر می‌شود.

۸- مراجع

[1] R.A.Kemmerer and G.Vigna, "Intrusion Detection:A Brief History and Overview," pp. 27-30, 2002.

[2] E. Beqiri, "Neural Networks for Intrusion Detection Systems," pp. 156–165, 2009.

[3] S. Zhao, Regina, and Saskatchewan, "Intrusion detection using the support vector machine enhanced with a feature-weight kernel," pp. 1-83, 2007.

[4] R. Richardson, "Computer Crime and Security Survey," pp. 1-30, 2007

[5] N. B. Idris and B. Shanmugam, "Artificial Intelligence Techniques Applied to Intrusion Detction," pp. 52-55, 2005.

[6] Prateek and D. S. K. Jena, "Intrusion Detection Using Self-Training Support Vector Machines," pp. 1-35, 2013.

[7] A. M. Chandrashekhar and K. Raghuveer, "Amalgamation of K-means Clustering Algorithm with Standard MLP and SVM Based Neural Networks to Implement Network Intrusion Detection System," vol. 2, pp. 273-283, 2014.

[8] W.C.Hong,"Traffic flowforecastingbyseasonalSVRwithchaoticsimulated annealing algorithm," vol. 74, pp. 2096–2107, 2011.

[9] V. N. Vapnik, "The Nature of Statistical Learning Theory," 1999.

[10] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers " pp. 144-152, 1992.

[11] V. N. Vapnik, "Statistical Learning Theory " 1998.

[12] D.Meyer, "Support Vector Machines," pp. 1-8, 2012.

[13] R. x. ZHANG, Z.r.DENG, and G. J. ZHI, "Intrusion Detection based on SVM and decision fusion," pp. 87-90, 2010.

[14] S. Mukkamala, G. Janoski, and A. Sung, "Intrusion Detection Using Neural Networks and Support Vector Machines," Department of Computer Science New Mexico Institute of Mining and Technology, pp. 1702-1707, 2002.

[15] S. J. Horng, M.Y.Su, Y. H. Chen, T. W. Kao, R. J. Chen, J. L. Lai, et al., "A novel intrusion detection system based on hierarchical clustering and support vector machines," Department of Computer Science and Information Engineering, pp. 306–313, 2011.

[16] Prateek and D. S. K. Jena, "Intrusion Detection Using Self-Training Support Vector Machines," pp. 1-35, 2013.

[17] S. R. Gunn, "Support Vector Machines for Classification and Regression," pp. 1-66, 1998.

[18] Y. Qian, H. Zhang, Y. Sang, and J. Liang, "Multigranulation decision-theoretic rough sets," International Journal of Approximate Reasoning, vol. 55, pp. 225–237, 2014.

[19] X. Wang, F. He, and L. Liu, "Application of Rough Set Theory to Intrusion Detection System," pp. 562-565, 2007.

[20] G. Kumar, K. Kumar, and M. Sachdeva, "The use of artificial intelligence based techniquesfor intrusion detection: a review," pp. 369–387, 2010.

[21] C. Bitter, J. North, D. Elizondo, and T. Watson, "An Introduction to the Use of Neural Networks for Network Intrusion Detection," pp. 5-24, 2012.

[22] N. B. Idris and B. Shanmugam, "Artificial Intelligence Techniques Applied to Intrusion Detection," pp. 52-55, 2005.

[23] J.P.Anderson, "Computer Security threat Monitoring And Surveillance," pp. 1-56, 1980

[24] G. Wang, J. Hao, J. Ma, and L.Huang, "A new approach to intrusion detection using Artificial Neural Networks andfuzzy clustering," pp. 6225–6232, 2010.

[25] S. R. Snapp, J. Brentano, T. L. G. G. V. Dias, C. L. H. L. T. Heberlein, K. N. Levitt, B. Mukherjee, S. E. Smaha1,, D. M. T. T. Grance, and D. Mansur4, "DIDS (Distributed Intrusion Detection System) - Motivation Architecture, and An Early Prototype ",pp. 167-176, 1991.

[26] [R. E. j. Kennedy" , Particle swarm optimization," in Neural Networks, Perth, WA 1995, pp. 1942 - 1948 .

[27] J. Chen, "Study on an Improved SVM Model for Intrusion Detection," pp. 275-280, 2012.

[28] S. Saha, A. S. Sairam, and A. Ekbal, "Genetic Algorithm Combined with Support Vector Machine for Building an Intrusion Detection System," pp. 566-572, 2012.

[29] J. T. Yao, S. Zhao, and L. Fan, "An Enhanced Support Vector Machine Model for Intrusion Detection," pp. 538-543, 2006.

[30] H. H. Gao, H. H. Yang, and X.Y.Wang, "Principal Component Neural Networks Based Intrusion Feature Extraction and Detection Using SVM," pp. 21-27, 2005.

[31] G. Xiaoqing, G. Hebin, and C. Luyi, "Network Intrusion Detection Method Based on Agent and SVM," pp. 1-4, 2010.

[32] M. K. Asif, T. A. Khan, and S. Yakoob, "Network Intrusion Detection and its Strategic Importance," pp. 140-144, 2013.

[33] S. A. Mulay, P. R. Devale, and G. V. Garje, "Intrusion Detection System using Support Vector Machineand Decision Tree," vol. 3, pp. 40-43, 2010.

[34] S. Mike, "Network Security Principles and Practices, Chapter 14," 2002

[۳۵] م. میرزائی و م. رحمانی, "تشخیص رخنه در شبکه‌های رایانه‌ای با رگرسیون بردار پشتیبان," نوآوری در مهندسی کامپیوتر و فناوری, تنکابن, ۱۳۹۲., ۷ صفحه.

[۳۶] ت. کریمی و م. صادقی مقدم, مجموعه‌های راف و مجموعه‌های خاکستری(مبانی، کاربرد، نرمافزار), تهران: موسسه کتاب مهربان نشر, تابستان ۱۳۹۳, ۱۶۶ صفحه.

[۳۷] م. میرزائی و م. رحمانی, "بررسی معماری‌های سرویس تشخیص رخنه," نوآوری در مهندسی کامپیوتر و فناوری, تنکابن, ۱۳۹۲, ۵ صفحه.

نرگس صالح‌پور، هم‌اکنون در دانشگاه علمی کاربردی لرستان فعالیت می‌کند. مدرک‌های کاردانی، کارشناسی و کارشناسی ارشد در رشتهٔ کامپیوتر (نرم‌افزار) را به‌ترتیب در دانشگاه علمی کاربردی لرستان، جهاد دانشگاهی لرستان و دانشگاه آزاد اسلامی واحد علوم تحقیقات لرستان، در سال‌های ۸۶، ۸۹ و ۹۳ دریافت کرده است. زمینهٔ پژوهشی وی، سامانه‌های تشخیص نفوذ با استفاده از تکنیک‌های داده‌کاوی است.

محمد نظری فرخی، مدرک‌های کارشناسی و کارشناسی ارشد خود را به‌ترتیب از دانشگاه پیام نور لرستان، ایران و دانشگاه آزاد اسلامی واحد علوم و تحقیقات لرستان در سال‌های ۸۹ و ۹۴ دریافت کرده است. زمینه پژوهشی وی سامانه‌های هوشمند است.

ابراهیم نظری فرخی، مدرک‌های کارشناسی،کارشناسی ارشد و دکترای خود را از دانشگاه آزاد اسلامی واحد علوم تحقیقات تهران، در سال‌های ۸۳، ۸۶ و ۹۳ دریافت کرده است. زمینه‌های پژوهشی وی سامانه‌های هوشمند و داده‌کاوی می‌باشد. همچنین ایشان عضو باشگاه پژوهشگران دانشگاه آزاد اسلامی واحد علوم تحقیقات می‌باشد و در حال حاضر نیز در دانشگاه تدریس می‌نماید.