Archive of SID

Quarterly Research Journal of Technical and Vocational University

Winter 2021, Vol. 17, No. 4, p. 13-33

Original Research

Detection of network penetration by data mining and using machine learning via SVM algorithm

Amir Abbas Namjooye Rad[1*], Mahdi Dadgarpour[2]

[1]Faculty Member, Department of Electrical and Computer Engineering, Faculty of Shahid Dadbin, Kerman branch, Technical and Vocational University (TVU), kerman, Iran.
[2]MA Student, Department of Information Technology, Faculty of E-Learning, Shiraz University, Shiraz, Iran.

ARTICLE INFO

Received: 01.26.2020
Revised: 12.11.2020
Accepted: 01.24.2021

Keyword:
SVM algorithm
data mining
machine learning
intrusion detection

***Corresponding Author:**
Amir Abbas Namjooye Rad
Email:
amir_namjoorad@yahoo.com

ABSTRACT

Computer networks are spreading widely and one of the most outstanding challenges in computer network security is detecting intrusions into networks. One of the main tools for detection is controlling network traffic and analyzing users' behavior. One way of accomplishing this is to set classifications that specify the patterns in huge volumes of data. By means of data mining methods and introducing a binary label (normal pack, abnormal pack) and specifying the priority of data, abnormal data is detected leading to increased accuracy of network intrusion detection which in turn leads to improvement and maintenance of network security. In this paper, SVM algorithm is analyzed in terms of priorities and the effect of machine learning algorithm on accuracy of intrusion detection is investigated. The results show that using SVM is more advantageous compared to past approaches yielding better detection and increasing accuracy and right alarm detection.

www.SID.ir

Archive of SID

فصلنامه علمی-پژوهشی کارافن

زمستان ۱۳۹۹، دوره ۱۷، شماره ٤، ۳۳-۱۳

مقاله پژوهشی

تشخیص نفوذ به شبکه به کمک داده‌کاوی و استفاده از یادگیری ماشین به روش ماشین بردار پشتیبان

امیرعباس نامجوی راد[۱*]، مهدی دادگرپور[۲]

۱- عضو هیئت علمی، دپارتمان مهندسی برق و کامپیوتر، آموزشکده شهید دادبین ، دانشگاه فنی حرفه ای استان کرمان، ایران.
۲- دانشجوی کارشناسی ارشد، دپارتمان مهندسی فناوری اطلاعات، دانشکده آموزش های الکترونیکی، دانشگاه شیراز، شیراز، ایران.

اطلاعات مقاله

دریافت مقاله: ۱۳۹۸/۱۱/۰۶
بازنگری مقاله: ۱۳۹۹/۰۹/۲۱
پذیرش مقاله: ۱۳۹۹/۱۱/۰۵

کلید واژگان:
سیستم تشخیص نفوذ
الگوریتم
یادگیری ماشین بردار پشتیبان
داده‌کاوی

***نویسنده مسئول:** امیرعباس نامجوی راد
پست الکترونیکی:
amir_namjoorad@yahoo.com

چکیده

با توجه به گسترش روزافزون شبکه‌های کامپیوتری، تشخیص نفوذ به شبکه، یکی از اجزای اصلی برقراری امنیت در شبکه‌های کامپیوتری شناخته می‌شود که ابزار اصلی آن، کنترل ترافیک شبکه و تحلیل رفتارهای کاربران است. یکی از راه‌های اجرای چنین سیستم‌هایی، استفاده از دسته‌بندی‌ها می‌باشد که با استفاده از مشخص کردن الگوها در حجم زیاد داده، کمک بزرگی به ما می‌کند. با استفاده از روش‌های داده‌کاوی و مشخص کردن یک برچسب دودویی (بسته نرمال، بسته غیرنرمال) و همچنین مشخص کردن ویژگی‌های داده‌ها که می‌توان داده‌های غیرنرمال را تشخیص داد؛ از این رو دقت درستی سیستم تشخیص نفوذ، افزایش می‌یابد و در نتیجه، امنیت شبکه بالا می‌رود. مدل پیشنهادی در این مقاله، به بررسی الگوریتم ماشین بردار پشتیبان در انتخاب خصیصه‌ها و تأثیر استفاده از الگوریتم‌های یادگیری ماشین در میزان دقت و میزان تشخیص نفوذ در سیستم می‌پردازد که نتایج حاصل نشان می‌دهد که استفاده از این الگوریتم، به افزایش میزان دقت و تشخیص درست هشدارها نسبت به روش‌های قبلی می‌انجامد.

www.SID.ir

Archive of SID

مقدمه

برای ایجاد امنیت کامل در یک سیستم کامپیوتری، علاوه بر دیواره‌های آتش و دیگر تجهیزات جلوگیری از نفوذ، سیستم‌های دیگری به نام سیستم‌های تشخیص نفوذ، نیاز می‌باشد تا بتوانند در صورتی که نفوذگر از دیواره آتش، آنتی‌ویروس و دیگر تجهیزات امنیتی عبور کرد و وارد سیستم شد، آن را تشخیص داد و راه‌حلی برای مقابله با آن پیدا کرد. سیستم‌های تشخیص نفوذ را می‌توان از سه جنبه معماری، نحوه پاسخ به نفوذ و روش تشخیص طبقه‌بندی کرد. انواع مختلفی از معماری سیستم‌های تشخیص نفوذ وجود دارد که به طور کلی می‌توان آنها را به سه دسته تقسیم‌بندی کرد که عبارتند از:

۱- مبتنی بر میزبان[1]: هر بسته ورودی به شبکه را برای حضور ناهنجاری‌های اطلاعات نادرست را کنترل می‌کند و براساس محتوا در هر آی‌پی یا سطح برنامه یک هشدار تولید می‌کند.

۲- مبتنی بر شبکه[2]: رفتار هر سیستم را تجزیه‌وتحلیل می‌کند، تفاوت امکانات NIDS بیشتر از HIDS است. این سیستم‌ها تنها از میزبان‌هایی که روی آنها مستقر هستند محافظت می‌کند.

۳- توزیع شده[3]: این سیستم از چندین NIDS یا HIDS یا ترکیبی از این دو نوع همراه یک ایستگاه مدیریت مرکزی تشکیل شده است. بدین صورت که هر IDS در شبکه موجود است گزارش‌های خود را برای ایستگاه مدیریت مرکزی ارسال می‌کند. ایستگاه مرکزی وظیفه بررسی گزارش‌های رسیده و آگاه‌سازی مسئول امنیتی سیستم را بر عهده دارد. این ایستگاه مرکزی همچنین وظیفه به‌روزرسانی پایگاه قوانین تشخیص هر یک از IDSهای موجود در شبکه را بر عهده دارد. اطلاعات در ایستگاه مدیریت مرکزی ذخیره می‌شود. شبکه بین NIDSها با سامانه مدیریت مرکزی می‌تواند خصوصی باشد یا اینکه از زیرساخت موجود برای ارسال داده‌ها استفاده می‌شود و وقتی از شبکه برای ارسال داده‌های مدیریتی استفاده شود امنیت‌های اضافی به وسیله رمزگذاری یا فناوری شبکه‌های خصوصی مجازی[4] حاصل می‌شود [۱]. الگوریتم‌های تشخیص نفوذ[5] از جنبه روش تشخیص، در دو دسته کلی طبقه‌بندی می‌شوند: تشخیص سوءاستفاده[6] و تشخیص ناهنجاری[7].

الگوریتم‌های تشخیص سوءاستفاده حمله‌ها را برمبنا امضای حملات شناخته‌شده شناسایی می‌کنند. با وجود اینکه این نوع از سیستم‌های تشخیص نفوذ در تشخیص حمله‌های شناخته شده با درصد خطای پایین کارآمدتر هستند اما نمی‌توانند حمله‌های جدیدی که ویژگی و خصوصیت مشابهی با حمله‌های شناخته‌شده ندارند را شناسایی کنند. در مقابل، الگوریتم‌های تشخیص ناهنجاری براساس این فرضیه هستند که رفتار حمله‌کننده با رفتار یک کاربر نرمال، متفاوت است؛ از این رو ترافیک‌های نرمال را آنالیز و الگوهای ترافیک نرمال را ایجاد می‌کنند. حال ترافیک ورودی را به‌عنوان حمله در نظر می‌گیرند اگر خصوصیات آن با الگوهای ترافیک نرمال متفاوت باشد. با وجود اینکه الگوریتم تشخیص ناهنجاری برای شناسایی حمله‌های جدید مناسب هستند اما در تشخیص حمله‌های شناخته‌شده به اندازه مدل‌های تشخیص سوءاستفاده، کارآمد نیستند[۲]. به دلیل حل معایب این دو روش تشخیص نفوذ معمولی، روش‌های تشخیص نفوذ ترکیبی از دو روش فوق ارائه شده است. در بسیاری از سیستم‌های تشخیص نفوذ ترکیبی یک مدل تشخیص سوءاستفاده و یک مدل تشخیص ناهنجاری به‌طور مستقل آموزش دیده و سپس نتایج آنها با هم جمع می‌شوند. برای مثال، بعضی از سیستم‌های تشخیص نفوذ ترکیبی، ترافیک ورودی را به‌عنوان حمله در نظر می‌گیرند اگر حداقل یکی از

[1] Host-based Detection System (HIDS)
[2] Network-based Intrusion Detection System (NIDS)
[3] Distributed Intrusion Detection System (DIDS)
[4] VPN
[5] Intrusion Detection Algorithm
[6] Misuse Detection
[7] Anomaly Detection

www.SID.ir

Archive of SID

دو مدل بیان‌شده تشخیص دهد که این ترافیک ارتباطی حمله است ترافیک ورودی را به عنوان حمله در نظر می‌گیرند و اگر هر دو مدل تشخیص دهند که ترافیک‌های ورودی حمله است ترافیک‌های ورودی را به‌عنوان حمله در نظر می‌گیریم. در حالت اول، سرعت کشف و شناسایی، بهبود می‌یابد اما سیستم تشخیص نفوذ هنوز نرخ مثبت کاذب[1] بالایی خواهد داشت در حالی که در حالت دوم، پیغام کاذب کاهش خواهد یافت اما ممکن است بسیاری از ترافیک‌های حمله را نادیده بگیرد [۳]. در این پژوهش به جای بررسی روش‌هایی که فقط نتایج دو مدل تشخیص نفوذ را ترکیب می‌کنند، روش‌های ترکیب سلسه‌مراتبی مورد مطالعه قرار می‌گیرد. لذا با این روش می‌خواهیم قبل از به‌کارگیری الگوریتم‌های کاهش ابعاد، در ابتدا بهترین الگوریتم را شناسایی می‌کنیم و الگوریتم‌هایی که از لحاظ کارایی نزدیک به هم هستند را ارزیابی و مقایسه خواهیم کرد. هدف این مقاله پیدا کردن دسته‌بندی با بیشترین نرخ دقت می‌باشد. همچنین این آزمایش‌ها در شبیه‌ساز و ابزار یادگیری ماشین وکا انجام می‌شود.

اهمیت و ضرورت تحقیق

هدف از این مقاله، استفاده از روش‌های مبتنی بر داده‌کاوی برای تشخیص نفوذ است؛ زیرا حملات همواره بروز می‌شوند و سیستم‌های تشخیص نفوذ ستنی نمی‌توانند این حملات شناسایی کنند. وقتی نفوذ اتفاق می‌افتد مهم‌ترین کار شناسایی است. رخداد مربوط به نفوذ در هر زمان مرتبط به الگویی ازاتفاقات است که در گذشته رخ داده است. این داده‌های تاریخی، منبع بسیار مهمی از صفات هستند که نیاز است تا به‌طور مؤثر علامت و نشانه‌های نفوذ در مجموعه داده‌ها مشخص شود. داده‌کاوی با کشف الگوهای مناسب از میان داده‌های قبلی، به روند ساخت این مدل‌ها کمک زیادی می‌کند. در این روش، مجموعه‌ای از قانون‌های دسته‌بندی از داده‌های شبکه به‌دست می‌آید. این قانون‌ها توانایی تعیین رفتار عادی از غیرعادی را دارا می‌باشند و با استفاده از مجموعه داده KDDcup99 ارزیابی می‌شوند. هدف اصلی ما معرفی بهترین الگوریتم با توجه به مجموعه داده‌ها است که بتواند بسته‌های عادی را از غیرعادی تشخیص دهد. نوآوری تحقیق استفاده از تمام الگوریتم‌های موجود در روش‌های دسته‌بندی است که در نرم افزار وکا موجود است و پیشنهاد پنج نمونه داده که از داده اولیه استخراج شده و برای مدل‌های مختلف و الگوریتم‌ها بهترین جواب را می‌دهد. استخراج پنج نمونه داده، زمان بسیار زیادی را به خود اختصاص می‌دهد و همه الگوریتم‌های مختلف موجود در مدل‌های دسته‌بندی با مجموعه داده‌های مختلف، شبیه‌سازی و اجرا شدند که در نهایت، پنج نمونه داده اولیه پیشنهاد شد.

فرضیه‌های تحقیق عبارتند از:

۱- می‌توان با طراحی یک سامانه ترکیبی در سیستم‌های تشخیص نفوذ علاوه بر داشتن مزایای سیستم‌های تشخیص نفوذ بر پایه تشخیص ناهنجاری به برخی از مزایای سیستم تشخیص سوءاستفاده رسید [۴].

۲- استفاده از روش تقسیم و غلبه در سیستم‌های تشخیص نفوذ باعث به دست آوردن کارایی بالاتر می‌گردد.

۳- یک پایگاه داده از حملات جهت آموزش سیستم طبقه‌بندی حملات وجود دارد.

اهداف این پژوهش عبارتند از:

۱- کارایی بیشتر در تشخیص نفوذ، در مقایسه با سیستم‌های دستی

۲- توانایی رسیدگی به حجم زیادی از اطلاعات

۳- توانایی هشدار نسبتاً بلادرنگ که باعث کاهش خسارت می‌شود.

۴- دادن پاسخ‌های خودکار، مانند قطع ارتباط کاربر، غیرفعال‌سازی حساب کاربر، اعمال مجموعه دستورهای خودکار و غیره

۵- افزایش میزان بازدارندگی

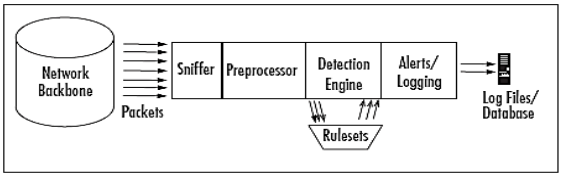
[1] False Positive

www.SID.ir

Archive of SID

۶- توانایی گزارش‌دهی.

برای ایجاد سیستم مدیریت هشدار پیشنهادی، ابزاری به صورت دقیق تعریف شده‌اند. به همین دلیل از ابزار اسنورت استفاده می‌شود. اسنورت، یک ابزار مهم تولیدکننده هشدار می‌باشد که معروف‌ترین سیستم تشخیص نفوذ کدباز است [۵ و۶] که قادر است دیتاگرام TCP/IP را روی شبکه به‌صورت هم‌زمان تحلیل کند. این ابزار، یک نرم‌افزار انعطاف‌پذیر است که می‌تواند به بیشتر پایگاه‌داده‌های مهم، مانند اوراکل[1] و ... متصل شود. اسنورت، یک موتور شناسایی نفوذ و یک پویشگر درگاه دارد که آن را قادر می‌سازد تا هر گونه حمله از پیش تعریف شده را شناسایی و پیام هشدار مناسب را تولید کند. اسنورت برای سیستم‌عامل‌های مختلف مانند لینوکس، یونیکس و ویندوز ویرایش‌های مختلف دارد [۷]. اسنورت، یک سیستم تشخیص نفوذ و جلوگیری از آن است، از یک زبان تعریف قانون استفاده می‌کند تا مزایای روش‌های شناسایی از طریق پروتکل، شناسایی از طریق امضا (نمونه‌های از پیش تعریف شده) و شناسایی از طریق رفتار غیرمعمول را با هم ادغام کند. مارتین روزچ؛ مؤسس شرکت source file، اسنورت را در سال‌های ۱۹۹۹ -۱۹۹۸ توسعه داد. این شرکت به پشتیبانی فنی از این ابزار می‌پردازد و بر اساس این ابزار برای سازمان‌های مختلف ساختارها و راه‌حل‌های امنیتی ارائه می‌دهد. اسنورت تا سال ۲۰۱۱ با بیش از ۳ میلیون دانلود متداول‌ترین فناوری تشخیص نفوذ در دنیا می باشد. اسنورت از چهار قسمت اصلی تشکیل شده است که در شکل ۱ مشاهده می‌شود.



شکل ۱. معماری اسنورت[۸]

قسمت جمع‌آوری اطلاعات

جمع‌کننده اطلاعات عبارت است از مجموعه‌ای از وسایل سخت‌افزاری و نرم‌افزاری که برای جمع‌آوری اطلاعات موردنیاز به شبکه موردنظر متصل شده‌اند. مشابه ابزاری که در سیستم تلفن به سیم تلفن وصل گردیده است، با این تفاوت که به جای صدای مکالمه در ارتباط تلفنی، اطلاعات شبکه کامپیوتری جمع‌آوری می‌گردد. در شبکه اینترنت معمولاً ترافیک IP قابل مشاهده می‌باشد در حالی که در شبکه‌های محلی ترافیک‌های دیگری از قبیل IPX و Apple Talk مشاهده می‌گردد.

- **قسمت آنالیز اولیه**

این قسمت وظیفه دسته‌بندی و آماده‌سازی اطلاعات برای مرحله بعد را به عهده دارد. در این قسمت، اطلاعات خام ورودی متناسب با محتوا دسته‌بندی می‌گردد. از جمله حالت‌های موجود می‌توان به سه حالت ذیل اشاره کرد:

- ✓ RPC plug-in
- ✓ HTTP plug-in
- ✓ Port scanner plug –in

این قسمت به‌صورت کمی در عملکرد اسنورت بسیار مؤثر است. این قسمت می‌تواند به‌صورت مستقل فعال یا غیرفعال شود؛ لذا قادر است بسیاری از کنترل‌های اولیه را انجام و نسبت به ارائه پیام، اقدام یا سکوت کند.

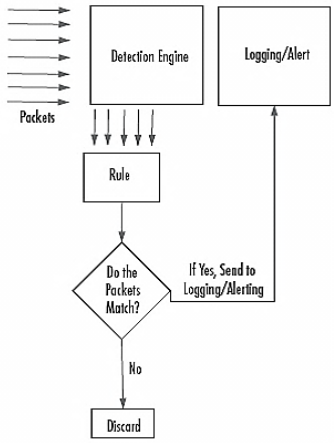
[1] MySQL Oracle

www.SID.ir

Archive of SID

- **موتور شناسایی**

پس از اینکه اطلاعات دسته‌بندی و آماده گردید تحویل این قسمت می‌گردد. قوانین (امضای حملات) در این قسمت قرار گرفته است. پس از ورود اطلاعات به این قسمت عملیات مقایسه و ارزیابی با قوانین انجام می‌گیرد. این قسمت دارای بیشترین تراکنش در عملیات تشخیص نفوذ می‌باشد. پس از مقایسه اطلاعات با قوانین موجود (امضای حملات)، در صورت تطابق با امضای حملات به قسمت هشدار ارسال می‌شود. در غیر این صورت اطلاعات رها و کنار گذاشته می‌شوند. در شکل ۲ روند کار موتور شناسایی نمایش داده شده است.



شکل ۲. موتور شناسایی[۸]

- **خروجی هشدار**

پس از تشخیص حمله، اطلاعات مورد نظر با فرمت از قبل پیش‌بینی شده به این قسمت ارسال می‌شود. با تنظیم اسنورت می‌توان اطلاعات را داخل پوشه لاگ در مجموعه دستورات اسنورت نگهداری ذخیره کرد. این اطلاعات با فرمت متن ذخیره می‌شود. در غیر این صورت می‌توان با کمک نرم‌افزارهای متعددی که دارای کنسول گرافیکی می‌باشند نسبت به جمع‌آوری اطلاعات خروجی اقدام نمود، از جمله این نرم‌افزارها می‌توان به موارد زیر اشاره کرد:

- ✓ SnortSnarf
- ✓ Snortplot.php
- ✓ Swatch
- ✓ ACID
- ✓ Kiwi Syslog Daemon

بعد دیگر از پیکربندی اسنورت به عنوان یک سیستم تشخیص نفوذ، استفاده از قوانین برای ایجاد معیار نفوذ برای اسنورت است. برای مثال می‌توان با قانونی، اسنورت را مکلف ساخت که نسبت به دسترسی‌های انجام‌شده مبتنی بر پروتکلی تعیین‌شده از/ به یک پورت خاص و از/ به یک مقصد معین با محتوایی شامل رشته‌ای خاص، اخطاری یا واکنشی ویژه را اعمال کند. اسنورت را می‌توان به گونه‌ای پیکربندی کرد که قابلیت تشخیص حمله توسط ابزارهای پویش پورت را نیز داشته باشد؛ لذا با وجود استفاده از اسنورت نیازی به استفاده از ابزاری ثانویه برای تشخیص پویش‌گرهای پورت وجود ندارد. اسنورت با قابلیت‌های نسبتاً کاملی که در خود جای داده‌است، به همراه رایگان بودن آن و قابلیت نصب بر محیط‌ها و سیستم‌های عامل متدوال، به یکی از معمول‌ترین IDSهای

www.SID.ir

Archive of SID

کنونی مبدل شده است. در ادامه، نحوه اعمال ترافیک KDDcup99 DARPA بر اسنورت برای تولید هشدار شرح می‌شود[۸].

مجموعه داده‌های KDDcup99

در هر IDS اولین قدم در داده‌کاوی، تأمین داده‌های ورودی است. این داده‌ها از منابع مختلف و با روش‌های مختلفی به‌دست می‌آیند. برای ردیابی رفتار غیرعادی در شبکه نت (نه در میزبان)، بهترین منبع ما ترافیک شبکه است که شامل بسته‌های فرستاده شده بین مبدأ و مقصد می‌باشد. در مجموعه داده‌هایKDDcup99، ترافیک شبکه‌ای میزبان یا شبکه نت جمع‌آوری می‌شود. در این مجموعه داده استاندارد ۴۹۴۰۲۰ نمونه و ۴۱ ویژگی ثبت شده است. خصیصه‌های مجموعه داده KDDcup99 به سه طبقه تقسیم می‌شوند:

۱- خصیصه‌های پایه: شامل تمام ویژگی‌هایی است که می‌توان از یک اتصال TCP/IP به‌دست آورد. بیشتر این ویژگی‌ها منجر به تأخیر در تشخیص می‌شوند.

۲- خصیصه‌های ترافیک: این طبقه شامل ویژگی‌هایی است که براساس فاکتورهای مشابه در یک بازه زمانی محاسبه می‌شوند که شامل دو گروه می‌شود:

الف) ویژگی‌های مشابه میزبان: این سیستم اتصالات ۲ ثانیه آخر را بررسی می‌کند. هر میزبان با مقصد مشابه یک اتصال جاری تلقی می‌شود و آمار مربوط به رفتارها، خدمات، پروتکل‌ها و غیره محاسبه می‌شود.

ب) ویژگی‌های مشابه سرویس: این ویژگی‌ها هم اتصالات سرویس یکسانی را دو ثانیه آخر اتفاق افتاده‌اند به‌عنوان اتصالات جاری در نظر می‌گیرند.

۳- خصیصه‌های محتوایی: برخلاف بیشتر حملات DOS و Probe حملات UZR و RZL نفوذها (هک‌های) تکراری و نمونه‌های متوالی ندارند که به این علت است که حملات DOS و Probe اتصالات زیادی با تعدادی میزبان در یک بازه زمانی کوتاه دارند. برای تشخیص چنین حملاتی تعدادی خصیصه در نظر گرفته شده‌اند (مثل ورود ناموفق) که قادرند رفتارهای مشکوک را تشخیص دهند. به این خصیصه‌ها، خصیصه‌های محتوایی می‌گویند. KDDcup99 از نظر طراحی و اجرای IDS یک مجموعه داده قابل‌اعتماد است. هر داده ثبت شده در این مجموعه دارای ۴۱ ویژگی است. معمولاً ویژگی‌ها سه شکل دارند شامل پیوسته، گسسته و نمادین که دارای دامنه ارزشی متفاوتی می‌باشند. در این مجموعه داده چهار حمله اصلی تشخیص داده شده، وجود دارد و خارج از این طبقه‌بندی‌ها برخی بسته‌ها ارزش نرمال (عادی) دارند و برخی دیگران حملات شناسایی نشده هستند.

اعمال ترافیک DARPA KDDcup99 بر اسنورت

DARPA98 دارای ۴ گیگابایت داده آموزشی دوتایی (دوگانه) خام فشرده است که از ۷ هفته ترافیک شبکه‌ای تشکیل شده است. داده‌های آزمایشی ۲ هفته حدود ۲ میلیون ثبت اتصال دارند. مجموعه داده آموزشی KDDcup99 شامل حدود ۵ میلیون ثبت اتصال است (یک اتصال یعنی توالی بسته‌های TCP که در زمان‌های مشخصی شروع و تمام می‌شوند که بین آنها داده‌ها از آدرس IP مبدأ به آدرس IP مقصد تحت یکسری پروتکل مشخص در جریان هستند) که هر کدام از آنها دارای ۴۱ ویژگی است و به‌عنوان متعارف یا حمله برچسب‌گذاری شده‌اند و هرکدام دارای یک نوع حمله مشخص هستند. حملات شبیه‌سازی‌شده به یکی از چهار گروه حمله سرویس (DOS)، حمله کاربر به روت (U2R)، حمله از راه دور به اتصال محلی (R2L) و حمله Probing تعلق دارند. ویژگی‌های KDDcup99 را می‌توان به سه گروه ویژگی‌های اصلی، محتوایی و ترافیک تقسیم کرد.

مجموعه داده، یکی از قابلیت‌های ابزار اسنورت خواندن ترافیک شبکه از فایل‌های دودویی ذخیره شده است. این فایل‌های دودویی شامل تمامی ترافیک شبکه در زمان مشخص می‌باشد. فایل‌های دودویی عبارت است از

www.SID.ir

Archive of SID

مجموعه‌داده جمع‌آوری‌شده قبلی و این مجموعه داده‌ها حاوی نمونه حمله‌های واقعی هستند استفاده از این مجموعه داده باعث صرفه‌جویی در زمان و هزینه می‌شود. همچنین امکان اعمال روش‌های مختلف روی همان مجموعه داده امکان‌پذیر است. برخی از این مجموعه‌داده‌ها DARPA، DEFCON [۹] هستند. روش پیشنهادی این مقاله از هشدارهای ایجاد شده از مجموعه‌داده DARPA وKDDcup99 استفاده می‌کند. مجموعه داده بیان شده شامل ترافیک شبکه مربوط به هفت هفته (که هر هفته شامل پنج روز است) می‌باشد. دو فایل از میان فایل‌های هر روز مجموعه‌داده، استفاده می‌شوند که عبارتند از outside.tcpdump و tcpdump.list. فایل outside.tcpdump همان فایل دودویی شامل ترافیک شبکه است و فایل tcpdump.list حاوی فهرست بسته‌های ارسالی در شبکه به همراه برچسب مربوط به بسته‌ها است. این برچسب مشخص می‌کند که آیا بسته مورد نظر بخشی از یک حمله خاص است یا ترافیک معمولی می‌باشد. قسمتی از فایل tcpdump.list در شکل ۳ نمایش داده شده است.

```
۱۷۸۹ ۰۶/۲۴/۱۹۹۹ 09:12:17 00:00:01 frag/i - - 207.103.080.104 172.016.118.020 1 pod
۱۷۹۰ ۰۶/۲۴/۱۹۹۹ ۰۹:۱۲:۱۷ ۰۰:۰۰:۰۱ frag/i - - ۲۰۷,۱۰۳,۰۸۰,۱۰٤ ۱۷۲,۰۱٦,۱۱۸,۰۲۰ 1 pod
۱۷۹۱ ۰۶/۲۴/۱۹۹۹ ۰۹:۱۲:۱۷ ۰۰:۰۰:۰۱ frag/i - - ۲۰۷,۱۰۳,۰۸۰,۱۰٤ ۱۷۲,۰۱٦,۱۱۸,۰۲۰ 1 pod
۱۷۹۲ ۰۶/۲۴/۱۹۹۹ ۰۹:۱۲:۱۷ ۰۰:۰۰:۰۱ frag/i - - ۲۰۷,۱۰۳,۰۸۰,۱۰٤ ۱۷۲,۰۱٦,۱۱۸,۰۲۰ 1 pod
٤۰٦٦ ۰۶/۲۴/۱۹۹۹ ۱۱:۳٦:۲۲ ۰۰:۰۰:۰۱ http ۲٥۱۹۰ ۸۰ ۱۷۲,۰۱٦,۱۱٥,۰۰٥ ۱۹۲,۱٥۰,۰۱۲,۱۰۳ 0 -
٤۰٦۷ ۰۶/۲۴/۱۹۹۹ ۱۱:۳٦:۲۷ ۰۰:۰۰:۰۱ http ۲٥۲٥۰ ۸۰ ۱۷۲,۰۱٦,۱۱٥,۰۰٥ ۱۳۰,۲۰۷,۲٤٤,۰۲٦ 0 -
٤۰٦۸ ۰۶/۲۴/۱۹۹۹ ۱۱:۳٦:۳۱ ۰۰:۰۰:۰۱ http ۲٥۲٥۱ ۸۰ ۱۷۲,۰۱٦,۱۱٤,۱٦۹ ۱۹۲,۱٥۱,۰۱۱,۰۳۲ 0 -
۱۳۸۱۷ ۰۶/۱۹/۱۹۹۹ ۱٤:۲۰:۰۲ ۰۰:۰۰:۰۱ eco/i - - ۲۰۸,۲٤۰,۱۲٤,۰۸۳ ۱۷۲,۰۱٦,۱۱۲,۱۸۰ 1 nmap
۱۳۸۱۸ ۰۶/۱۹/۱۹۹۹ ۱٤:۲۰:۰۲ ۰۰:۰۰:۰۱ eco/i - - ۲۰۸,۲٤۰,۱۲٤,۰۸۳ ۱۷۲,۰۱٦,۱۱۲,۱۸۱ 1 nmap
۱۳۸۱۹ ۰۶/۱۹/۱۹۹۹ ۱٤:۲۰:۰۲ ۰۰:۰۰:۰۱ eco/i - - ۲۰۸,۲٤۰,۱۲٤,۰۸۳ ۱۷۲,۰۱٦,۱۱۲,۱۸۲ 1 nmap
```

شکل ۳. اعمال ترافیک DARPA KDDcup99

همان‌گونه که مشاهده می‌شود سه برچسب pod، nmap و به ترتیب مربوط به حمله pod، nmap و ترافیک معمولی می‌باشد.

داده‌کاوی

داده‌کاوی، پل ارتباطی میان علم آمار، علم کامپیوتر، هوش مصنوعی، الگوشناسی، فراگیری ماشین داده می‌باشد. داده‌کاوی، فرایند استخراج اطلاعات معتبر، از پیش ناشناخته، قابل‌فهم و قابل‌اعتماد از پایگاه داده‌های بزرگ و استفاده از آن در تصمیم‌گیری در فعالیت‌های تجاری مهم است. اصطلاح داده‌کاوی، به فرایند نیم‌خودکار تجزیه و تحلیل پایگاه داده‌های بزرگ به منظور یافتن الگوهای مفید اطلاق می‌شود. داده‌کاوی، فرایندی پیچیده برای شناسایی الگوها و مدل‌های صحیح، جدید و به‌صورت بالقوه مفید، در حجم وسیعی از داده می‌باشد، به طریقی که این الگوها و مدل‌ها برای انسان‌ها قابل درک باشند. داده‌ها اغلب حجیم می‌باشند و به‌تنهایی قابل استفاده نیستند، اما دانش نهفته در داده‌ها قابل استفاده می‌باشد. بنابراین بهره‌گیری از قدرت فرایند داده‌کاوی برای شناسایی الگوها و مدل‌ها و نیز ارتباط عناصر مختلف در پایگاه داده به‌منظور کشف دانش نهفته در داده‌ها و نهایتاً تبدیل داده به اطلاعات، روزبه‌روز ضروری‌تر می‌شود. در داده‌کاوی معمولاً به کشف الگوهای مفید از میان داده‌ها اشاره می‌شود. منظور از الگوی مفید، مدلی در داده‌ها است که ارتباط میان یک زیرمجموعه از داده‌ها را توصیف می‌کند و معتبر، ساده، قابل‌فهم و جدید است[۱۰].

www.SID.ir

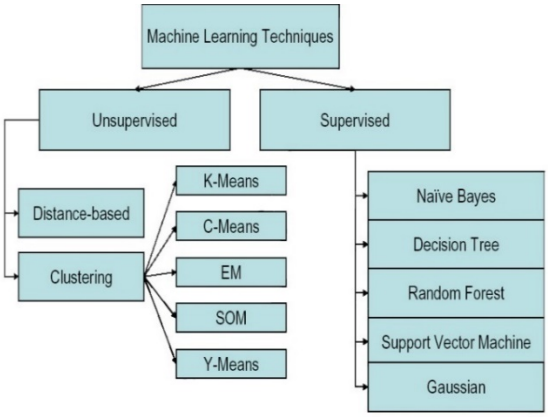
Archive of SID

نرم‌افزار WEKA

نرم‌افزار وکابا داشتن امکانات بسیار گسترده، امکان مقایسه خروجی روش‌های مختلف با هم، راهنمای خوب، واسط گرافیگی کارا، سازگاری با سایر برنامه‌های ویندوزی معرفی شد. میزکاروکا، مجموع‌های از الگوریتم‌های روز یادگیری ماشینی و ابزارهای پیش‌پردازش داده‌ها می‌باشد. این نرم‌افزار به گونه‌ای طراحی شده است که می‌توان به سرعت، روش‌های موجود را به صورت انعطاف‌پذیری روی مجموعه‌های جدید داده، آزمایش کرد. این نرم‌افزار، پشتیبانی‌های ارزشمندی را برای کل فرایند داده‌کاوی‌های تجربی فراهم می‌کند. این پشتیبانی‌ها، آماده‌سازی داده‌های ورودی، ارزیابی آماری چارچوب‌های یادگیری و نمایش گرافیکی داده‌های ورودی و نتایج یادگیری را در برمی‌گیرند. همچنین، هماهنگ با دامنه وسیع الگوریتم‌های یادگیری، این نرم‌افزار شامل ابزارهای متنوع پیش‌پردازش داده‌هاست. این جعبه ابزار متنوع و جامع، از طریق یک واسط متداول در دسترس است، به نحوی که کاربر می‌تواند روش‌های متفاوت را در آن با یکدیگر مقایسه کند و روش‌هایی را که برای مسائل مدنظر مناسب‌تر هستند، تشخیص دهد.

یادگیری ماشین

یادگیری ماشین زمینه نسبتاً جدیدی از هوش مصنوعی است که در حال حاضر دوران رشد و تکامل خود را می‌گذراند یادگیری ماشین یک زمینه تحقیقاتی بسیار فعال در علوم کامپیوتر است. علوم مختلفی از قبیل هوش مصنوعی، روانشناسی، فلسفه، تئوری اطلاعات، آمار و اطلاعات، تئوری کنترل با یادگیری ماشین در ازتباط هستند. یادگیری ماشین عبارت است از اینکه چگونه می‌توان رنامه‌ای نوشت که از طریق تجربه یادگیری کرده و عملکرد خود را بهتر کند در یادگیری ماشین با استفاده از تئوری اطلاعات مدل‌های ریاضی ساخته می‌شود که می‌تواند برای استنتاج استفاده شوند. تکنیک‌های یادگیری ماشین در شرایطی مناسب است که هیچ گونه دانش اولیه در مورد الگوهای داده‌ها وجود ندارد؛ به همین دلیل گاهی به این روش‌ها پایین به بالا می‌گویند. مزیت مهم این روش این است که معمولاً به انسان‌های خبره برای تعیین ملزومات موردنظر به منظور تشخیص نفوذ نیازی نیست به همین دلیل بسیار سریع عمل می‌کنند و مقرون‌به‌صرفه هستند. تکنیک‌های یادگیری ماشین در داده‌کاوی به طور کلی به دو دسته نظارت‌شده و غیرنظارت‌شده تقسیم‌بندی می‌شوند. شکل ۴ روش‌های یادگیری ماشین و زیرمجموعه‌های موجود در آنها را نمایش می‌دهد که در ادامه قابل مشاهده است.



شکل ۴. تکنیک‌های یادگیری ماشین

www.SID.ir

Archive of SID

یادگیری انواع مختلف دارد که به شرح زیر هستند:
کلاس‌بندی[1]: ماشین یاد می‌گیرید ورودی‌ها را به دسته‌های از پیش تعیین شده نسبت دهد.
خوشه‌بندی[2]: سیستم یادگیر کشف می‌کند که کدام ورودی‌ها با هم در یک دسته قرار می‌گیرند.
تخمین عددی[3]: ماشین یاد می‌گیرد به جای تعیین دسته‌بندی یک ورودی مقدار عددی آن را پیش‌بینی نماید.

الگوریتم ماشین بردار پشتیبان

الگوریتم ماشین بردار پشتیبان، در هر جایی که نیاز به تشخیص الگو یا دسته‌بندی اشیا در کلاس‌های خاص باشد استفاده می‌شود. ماشین بردار پشتیبان، دسته‌بندی‌کننده‌ای است که جزو شاخه روش هسته[4] دریادگیری ماشین محسوب می‌شود. ماشین بردار پشتیبان را واپنیک در سال ۱۹۹۲ معرفی کرد و بر پایه تئوری یادگیری استاتیک[5] بنا گردیده است. هدف این دسته الگوریتم‌ها تشخیص و متمایز کردن الگوهای پیچیده از طریق کلاسترینگ، دسته‌بندی، رنکینگ، پاکسازی و غیره در داده‌ها است. ایده اصلی در ماشین بردار پشتیبان این است که:

- با فرض اینکه دسته‌ها به‌صورت خطی جداپذیر باشند، ابرصفحه‌هایی با حداکثر حاشیه[6] را به‌دست می‌آورد که دسته‌ها را جدا کنند.
- در مسایلی که داده‌ها به‌صورت خطی جداپذیر نباشند داده‌ها به فضای با ابعاد بیشتر نگاشت پیدا می‌کنند تا بتوان آنها را در این فضای جدید به‌صورت خطی جدا کرد.
- در مورد ماشین بردار پشتیبان تحقیقات زیادی در کشور انجام شده است؛ مانند صالح‌پور و همکاران (۱۳۹۳) نشان دادند که این الگوریتم علی‌رغم دقت بالایی که دارد اما بهتر است که در کاربردهای عملی به‌عنوان ورودی شبکه قرار گیرد[۱۱].

ماشین بردار پشتیبان یک متد برمبنای یادگیری ماشین است که اساس آن یادگیری با کمک اطلاعات است که داده‌ها را با کمک بردارهای پشتیبان که الگوهای داده‌ای را بیان می‌کنند کلاس‌بندی می‌کند. برای دسته‌بندی داده‌ها به دو دسته باید تابع $f(X)$ پیدا کنیم؛ به طوری که $Y_i = f(x_i)$ برای N داده تابع f را بتوان به این صورت تعریف کرد:

$$F(x) = \mathrm{sgn}(\sum_{i=1}^{i} a_i y_i (x_i \cdot x - b)) \qquad (۱)$$

که در آن ۱ تعداد رکوردهایی هستند که برای آموزش استفاده می‌شوند $a_i y_i \in \{-1,1\}$ عددی مثبت و کوچک‌تر از عدد ثابت C است و x_i نیز بردار پشتیبان است. اما اگر تابعی که برای دسته‌بندی استفاده می‌شود خطی نشود، باید داده‌ها را به ابعادی بالاتر ببریم تا تابع جداکننده آنها تابعی خطی بشود. برای این کار تبدیل به صورت زیر درمی‌آید که در آن $k(x_i,x)$ تابع هسته نامیده می‌شود که همان تابعی است که برای بردن داده‌ها به بعد بالاتراستفاده می‌شود.

$$F(x) = \mathrm{sgn}(\sum_{i=1}^{i} a_i y_i k(x_i,x) + b) \qquad (۲)$$

[1] Classification
[2] Clustering
[3] Numeric predication
[4] Kernel Methods
[5] statistical learning theory
[6] maximum margin

www.SID.ir

Archive of SID

تابع هسته می‌تواند انواع مختلفی داشته باشد و به دسته‌های مختلفی تقسیم شود که به علت پیچیدگی ریاضی بیان آنها مشکل است. بیشتر انواع توابع هسته به صورت خطی هستند و تفاوتی میان خصوصیت‌های مختلف داده‌ها قائل نمی‌شوند. تابعی که در قسمت قبلی بیان شد یک تابع خطی است که همین خصوصیت را دارد که در آن با همه خصوصیات داده‌ها به طور یکسان رفتار می‌شود. این یکسان بودن رفتار کارایی را پایین می‌آورد و بر دقت ماشین بردار پشتیبان اثر منفی می‌گذارد. راه‌حلی که برای این کار به نظر می‌رسد این است که برای تابع هسته وزن در نظر بگیریم. این وزن‌ها برای تعیین اثر خصوصیت‌ها به کار می‌روند. در کاربرد مدرن یادگیری ماشینی، ماشین بردار پشتیبان یکی از قوی‌ترین و بادقت‌ترین روش‌ها در الگوریتم‌های یادگیری ماشینی است. ماشین بردار پشتیبان یکی از متدهای یادگیری امتحان شده است که برای طبقه‌بندی، پیش‌بینی و رگرسیون استفاده می‌شود. این متد، روشی نسبتاً جدید است که در سال‌های اخیر عملکرد خوبی در طبقه‌بندی در مقایسه با متدهای قدیمی‌تر مثل شبکه‌های عصبی پرسپترون[1] داشته و روش بسیار ساده‌ای نیز می‌باشد. ماشین بردار پشتیبان به علت توانایی بالایش در طبقه‌بندی و برتری‌اش نسبت به سایر الگوریتم‌های طبقه‌بندی و رگرسیون بسیار محبوب شده است. در واقع ماشین بردار پشتیبان برای طبقه‌بندی دوتایی[2] طراحی شده است. بنابراین فرایند آن به سمت حل مسایل کلاسه‌بندی موجود بین رفتار عادی و غیرعادی یا مشکوک، بسیار مفید است. ماشین بردار، پشتیبان یک یادگیری ماشینی است که بردارهای یادگیری را در فضاهایی با ابعاد ویژگی‌های بالا ترسیم می‌کند و به هر بردار طبقه متناظرش برچسب می‌زند. اساس کار خوشه‌بند ماشین بردار پشتیبان خوشه‌بندی داده‌ای خطی است و در بخش خطی داده‌ها، مهم‌ترین نکته این است که خط انتخاب شده بیشترین حاشیه اعتماد را داشته باشد. یافتن بهترین معادله خطی برای داده‌ها از طریق روش‌هایQP انجام می‌شود که روش‌های معروفی در حل مسایل دارای محدودیت می‌باشند. قبل از عملیات تقسیم‌بندی خطی برای توانمندسازی ماشین برای طبقه‌بندی اطلاعات پیچیده‌تر، داده‌ها از طریق عملیات Phi به فضایی با بعد بالاتر منتقل می‌شوند. مشکلات به‌وجود آمده در ابعاد بالاتر را می‌توان از طریق قضیه دوتایی لاگرانژ برای تبدیل مسئله کمینه موردنظر به شکل دوتایی حل کرد. بدین روش به جای استفاده از روش پیچیده‌ای مثل Phi می‌توان از عملیات ساده‌تری مثل کرنل استفاده کرد که بردار مضربی از Phi می‌باشد. فرایند توسعه مدل ماشین بردار پشتیبان طبق آنچه در زیر می‌آید شامل سه مرحله است: پیش‌پردازش داده‌ها، توسعه مدل، استخراج مدل و در نهایت اجرا کردن آن. این الگوریتم قابل مقیاس‌گذاری است و حجم داده‌ها در عملکرد آن تأثیری ندارد. بنابراین پیچیدگی طبقه‌بندی بستگی به ابعاد فضای ویژگی‌ها ندارد. لذا در مقایسه با شبکه‌های عصبی، از این روش می‌توان برای یادگیری مجموعه بزرگی از مدل‌ها استفاده کرد. این الگوریتم براساس ساختار به حداقل‌رسانی ریسک طراحی شده و کاربردهای زیادی دارد و بسیار موفق بوده است[۱۲].

طراحی روش پیشنهادی

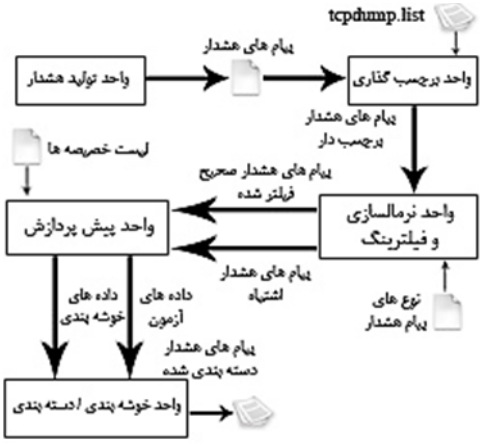
در این بخش، سیستمی برای انتخاب خصیصه با استفاه از الگوریتم ماشین بردار پشتیبان در واحد پیش‌پردازش سیستم تشخیص نفوذ ارائه می‌شود. واحد پیش‌پردازش، وظیفه تبدیل مقادیر مختلف خصیصه‌ها به مقادیر قابل قبول سیستم را دارد که در این پژوهش به خصیصه‌های موردنظر اعدادی نسبت داده شده است. همچنین در ادامه، عمل خوشه‌بندی توسط الگوریتم درخت ID3 ارائه شده است. شکل ۵-۳ شمای کلی این سیستم را نشان می‌دهد. این ساختار از چندین واحد و تعدادی فایل تشکیل می‌شود که به عنوان ورودی یا خروجی این واحدها هستند. سیستم ارائه شده مستقیماً به هشدارهای تولید شده توسط سیستم تشخیص نفوذ وابسته می‌باشد. این

[1] Perceptron neural networks
[2] Binary

www.SID.ir

Archive of SID

امر بدین معنی است که سیستم، هشدارهای تولید شده توسط سیستم تشخیص نفوذ را به عنوان ورودی گرفته و سپس شروع به پردازش آن‌ها می‌کند. برای آماده‌سازی هشدارها از ابزار اسنورت [۱۳] به همراه مجموعه داده DARPA KDDcup99 استفاده می‌شود. ابزار اسنورت ترافیک مستقیم DARPA KDD cup 99 را دریافت و سپس فایل‌های ثبت هشدار را ایجاد می‌کند. از این فایل‌های ثبت هشدار، به‌عنوان ورودی‌های اولیه سیستم استفاده می‌شود [۱۴].



شکل ۵. شمای کلی سیستم خوشه‌بندی و دسته‌بندی هشدار[۱۳]

در شکل ۵ روند اجرای خوشه‌بندی در روش پیشنهادی به صورت شماتیکی نمایش داده شده است. در این شکل، مراحل انجام خوشه‌بندی و دسته‌بندی هشدار توصیف شده است.

اجرا

در روش پیشنهادی از الگوریتم ماشین بردار پشتیبان برای انتخاب خصیصه‌های مورد نظر استفاده شده است روش تقسیم‌بندی داده‌ها با استفاده از ماشین بردار پشتیبان از روش‌های شناخته شده در کلاس‌بندی داده‌ها می‌باشد که نیازی به تنظیم پارامترها از قبل و همچنین دانش اولیه از داده‌ها نمی‌باشد. مجموعه داده‌های آموزشی D شامل n عضو(نقطه) را در اختیار داریم که به صورت زیر تعریف می‌شود:

$$D = \{(x_i,y_i) \mid x_i \in R^p, y_i \in \{-1,1\}\}_{i=1}^{n} \quad (۳)$$

مقدار y_i برابر ۱ یا 1-و هر x_i یک بردار حقیقی p- بعدی است. هدف پیدا کردن ابرصفحه جداکننده با بیشترین فاصله از نقاط حاشیه‌ای است که نقاط با $y_i = 1$ را از نقاط با $y_i = -1$ جدا کند. هر ابر صفحه می‌تواند به صورت مجموعه‌ای از نقاط x که شرط زیر را برآورده می‌کند نوشت:

$$w.x-b=0 \quad (۴)$$

$$\text{if } Y_i = +1 \qquad wx_i + b \geq 1 \quad (۵)$$

$$\text{if } Y_i = -1 \qquad wx_i + b \leq 1 \quad (۶)$$

$$\text{for all } i \quad y_i(w_i + b) \geq 1 \quad (۷)$$

www.SID.ir

Archive of SID

در این معادله x یک نقطه برداری و w یک وزن برداری است. بنابراین برای جداسازی داده‌ها همیشه باید بیشتر از ۰ باشد. در میان تمام ابرصفحه‌های بالقوه، ماشین بردار پشتیبان را انتخاب می‌کند که فاصله ابرصفحه آن حداکثر ممکن باشد. اگر داده‌های آزمون منطقی باشد و هر بردار آزمون در شعاع r از بردار آموزش قرار داشته باشد. اگر ابرصفحه انتخابی، در بیشترین فاصله از داده‌ها قرار گرفته باشد، حاشیه‌ها را به حداکثر می‌رساند و خطوط نزدیک‌ترین نقاط مجموعه داده‌های پوسته محدب دو مجموعه نصف می‌شود. فاصله نزدیک‌ترین نقطه روی ابرصفحه به مبدا با به حداکثر رساندن x به‌دست می‌آید چون x روی ابرصفحه است. برای نقاط سمت دیگر نیز استراتژی مشابهی را در پیش می‌گیریم. با ثابت‌سازی و کسر دو فاصله، مجموع فاصله ابرصفحه جداکننده از نزدیک‌ترین نقاط به‌دست میآید.

برای تمام نقاط $\{(X_i)\}: y_i(w_t x_i + b) \geq 1$، باید منشور قایم درجه دوم را برای تمام محدوده‌های خطی بهینه کرد. راه‌حل شامل ساختن یک مسئله دو مجهوله است که در آن ضریب لاگرانژ a_i با تمام محدوده‌ها در مسئله اولیه در ارتباط است که به حداکثر رسیده و برای تمام a_i، $a_i \geq 0$ محاسبه می‌شود.

راه‌حل به شکل زیر است: $Q(a) = \sum a_i - \frac{1}{2}\sum\sum a_i a_j y_i y_j x_i T x_j$

$$w = \sum a_i y_i x_i \quad (۸)$$

$$b = y_k - w_t x_k \quad (۹)$$

حتی a_iهای غیر صفر که معادل xi هستند یک محور پشتیبان می‌باشند. پس طبقه‌بندی شکل زیر را صورت خواهد گرفت:

$$f(x) = \sum a_i y_i x_i t_x + b \quad (۱۰)$$

که بستگی به یک محصول داخلی بین نقطه آزمون x و بردارهای پشتیبان xi دارد. حل مسئله بهینه‌سازی شامل محاسبه محصولات داخلی $x_i T x_j$ بین تمام جفت نقطه‌های اموزش می‌باشد. فرمول قدیمی: پیدا کنید wو b را به‌طوری که $\Phi(w) = ½\ wTw$ به حداقل برسد و برای تمام $\{(x_i,y_i)\}, y_i(wTx_i + b) \geq 1$ در ادامه الگوریتم ماشین بردار پشتیبان برای تشخیص نفوذ با توجه به TrainDوTestD توضیح داده می‌شود.

$$MM = \frac{2}{\|W\|} \quad (۱۱)$$

$$\Phi(W) \frac{1}{2} W^t W \quad (۱۲)$$

$$f(x) = \sum a_i y_i x_i^T x + b \quad (۱۳)$$

پیش‌پردازش، اولین و مهم‌ترین گام در استفاده از ابزار استخراج داده می‌باشد. برای رسیدن به درست‌ترین نتایج از الگوریتم‌های استخراج داده، فرایند کشف اطلاعات نهفته در مجموعه داده باید انجام شود. در این الگوریتم داده‌ها باید مجزا باشند تا بتوان آنها را وارد الگوریتم lib ماشین بردار پشتیبان کرد. برای اجرای ماشین بردار پشتیبان هر نقطه داده باید توسط یک بردار شامل اعداد حقیقی نشان داده شود. برای یک ویژگی با تعداد m ردیف، m تعداد عدد استفاده می‌شوند و فقط یکی از این اعداد برابر با ۱ و بقیه برابر با ۰ هستند. بنابراین الگوریتم

www.SID.ir

Archive of SID

Discretize استفاده شده است. اندازه‌گیری داده‌ها قبل از اجرای ماشین بردار پشتیبان آن‌قدر مهم است که ویژگی‌های با بازه ارزشی کوچک‌تر توسط ویژگی‌هایی با بازه ارزشی بزرگ‌تر تحت تأثیر قرار نگیرند. مزیت دیگر این روش اجتناب از روبه‌رو شدن با مشکلات عددی در طول محاسبات است. از آنجایی که مقادیر هسته‌ای معمولاً مربوط به محصول درونی بردارهای ویژگی‌ها هستند (مثلاً هسته خطی یا هسته چند فرمولی) مقادیر بزرگ ویژگی‌ها ممکن است باعث ایجاد مشکلات عددی شوند. برای حل این مشکل هر ویژگی باید به‌صورت خطی روی نقاط [1, -1] یا [1, 0] اندازه‌گیری شود. گام بعدی، انتخاب ویژگی است. الگوریتم‌های زیادی برای انتخاب ویژگی وجود دارند. در این پروژه از ماشین بردار پشتیبان برای انتخاب استفاده شده است. بعد از تنظیم پارامترهای موجود در ماشین بردار پشتیبان ویژگی‌هایی که بالاترین تأثیر را در عملکرد سیستم دارند برای آزمایش دقت تشخیص استفاده می‌شوند. در این الگوریتم هسته RBF استفاده شده است. هسته RBF نمونه‌ها را به‌صورت غیرخطی وارد فضایی با ابعاد بزرگ‌تر می‌کند.RBF مشکلات عددی کمتری ایجاد می‌کند. در این روش، داده‌های آموزشی به دو بخش برابر تقسیم می‌شوند که یکی از آنها مبهم فرض می‌شود. بنابراین پیش‌بینی صحت مربوط به این مجموعه داده‌ها می‌تواند عملیات طبقه‌بندی را در مواجهه با داده‌های مبهم نشان دهد. در نرم‌افزار وکا تابع Lib ماشین بردار پشتیبان برای به‌کارگیری الگوریتم ماشین بردار پشتیبان استفاده شده است. این تابع یک مجموعه کامل برای طبقه‌بندی ماشین بردار پشتیبان (nu_svc)، رگرسیون (epsilon_svr, nu_svr) و برآورد توزیع شده است. این تابع برای پشتیبانی از طبقه‌بندی چند مرحله‌ای طراحی شده است.

الگوریتم ماشین بردار پشتیبان مورد استفاده

داده‌های ورودی: مجموعه داده آموزش TrainD، مجموعه داده آزمون TestD که طبقه‌بندی نشده است.

داده‌های خروجی: مجموعه داده آزمون TestD که طبقه‌بندی شده است.

۱- تمام داده مجموعه داده آموزش را با (Xi, Yj) شروع کنید که در آن X بردار داده x1, ……, xn و Y بردار طبقات است.

۲- بردار وزن W را شروع کنید.

۳- تمام نقاط (x, y) را پخش کنید و جداکننده ابرصفحه را استخراج کنید.

۴- اگر ابرصفحه جداسازی بهینه داشته باشد از آن برای طبقه‌بندی مجموعه داده TestD استفاده کنید و پروسه را به پایان برسانید. در غیر این صورت گام‌های زیر را ادامه دهید.

۵- از طریق معادله (۳-۱۷) ابرصفحه را ماکزیمم و از طریق معادله (۳-۱۸) آن را مینیمم کنید.

۶- از طریق معادله ضریب لاگرانژ αi، بردار α1…αn را شروع کنید.

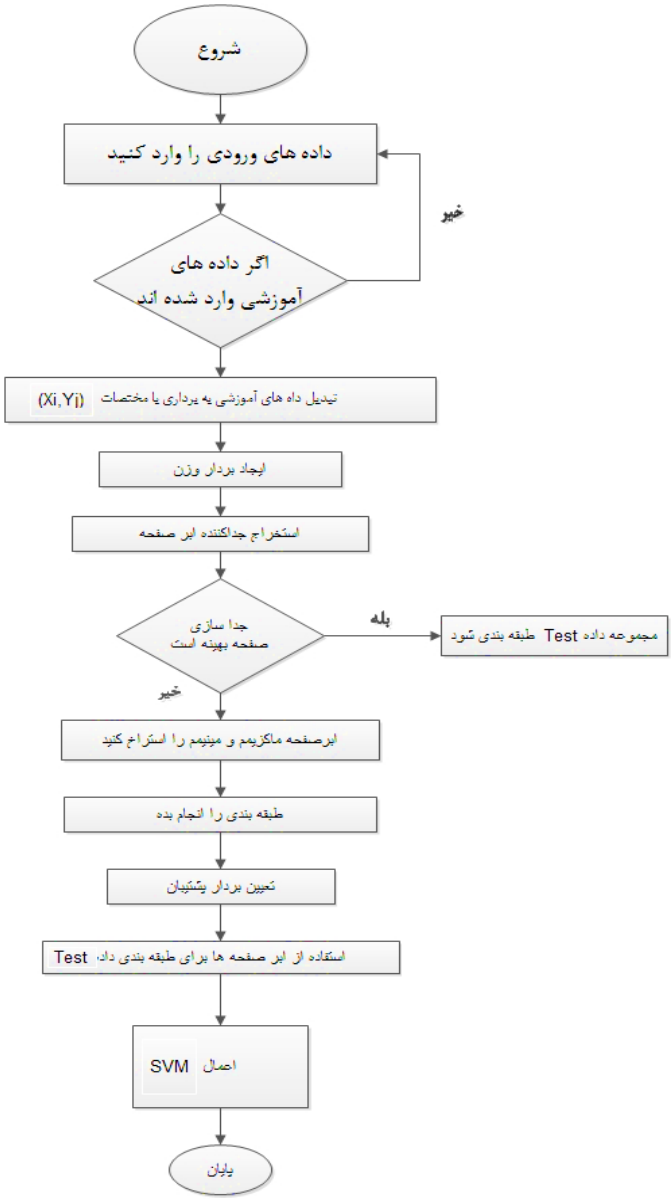
۷- از طریق معادله ۱۲ طبقه‌بندی را انجام دهید.

۸- بردارهای پشتیبان xi را با αi ناصفر تعیین کنید (یعنی نقاطی که محدوده ابرصفحه را نشان می‌دهند.)

۹- از ابرصفحه به‌وجودآمده از تعیین بردارهای پشتیبان به‌عنوان مدل طبقه‌بندی مجموعه داده آزمون TestD استفاده کنید. yi(w·xi−b)≥1.{\displaystyle y_{i}(\mathbf {w} \cdot \mathbf {x_{i}} -b)\geq 1.\,} 1. با اعمال الگوریتم ماشین بردار پشتیبان روی فایل ورودی نتایجی که در ادامه ارائه خواهند شد به‌دست می‌آیند.

www.SID.ir

Archive of SID



شکل ۶. روند نمای روش پیشنهادی

نتایج به دست آمده

برای شبیه‌سازی سیستم ارائه‌شده، از نرم‌افزار وکا استفاده شده است. چندین ابزار جانبی توسط نرم‌افزار متلب نیز اجرا گردید که عملیات برچسب‌گذاری هشدارها، پیش‌پردازش آنها و عملیات ترکیب خوشه‌ها را انجام می‌دهند.

www.SID.ir

Archive of SID

برای شبیه‌سازی سیستم ارائه شده ابتدا باید فایل‌های هشدار تولید شوند. برای این منظور ابزار اسنورت نصب می‌شود و سپس فایل‌های دودویی KDDcup99،DARPA که حاوی ترافیک ذخیره شده می‌باشند برای بررسی به اسنورت داده می‌شوند. ابزار اسنورت، این فایل‌ها را بررسی و هشدارهایی را تولید می‌کند. نرم‌افزار برچسب‌گذاری هشدارها، فهرست هشدارها را به همراه فایل‌های هشدار تولید شده، به‌عنوان ورودی دریافت و هشدارهای برچسب‌گذاری‌شده ایجاد می‌کند. برچسب‌های ایجاد شده برای هشدارها برای آموزش و آزمون صحت عملکرد و ایجاد داده‌های آماری استفاده می‌شوند. پس از برچسب‌گذاری هشدارها، هشدارها وارد واحد فیلتر کردن و نرمال‌سازی می‌شوند. نحوه عملکرد این واحد در بخش‌های قبلی توضیح داده شده است. خروجی این بخش دو دسته می‌باشد که یکی حاوی هشدارهای اشتباه و دیگری حاوی هشدارهای صحیح برچسب‌دار است. بعد از این مرحله این دو دسته هشدار وارد واحد پیش‌پردازش می‌شوند. چون سیستم تشخیص نفوذ برای کار خود نیاز به بردارهای داده‌ای دارد؛ در این بخش، هشدارهای وارد شده به بردارهای داده‌ای تبدیل می‌شوند. پس از تبدیلات لازم و تولید بردارهای داده‌ای، این بردارها وارد بخش خوشه‌بندی و دسته‌بندی می‌شوند و ابتدا توسط داده‌های آموزش، آموزش می‌بینند. در ادامه، الگوریتم پیشنهادی در محیط وکا و با دیتاست‌های معرفی شده ارزیابی شد. نتایج به‌دست‌آمده از مقایسه‌ای که با روش‌های پیشین انجام شده است حاکی از این قضیه است که دقت شناسایی نسبت به روش‌های پیشین، بهبود یافته است.

در ارزیابی نتایج به‌دست‌آمده، از عباراتی مانند میزان دقت و درستی و ... استفاده می‌شود که در این قسمت به ارائه روابط و توصیف هر یک پرداخته شده است. معیارهای ارزیابی عملکرد الگوریتم‌های یادگیری معمولاً شامل درستی[1]، دقت[2]، صحت[3] می‌باشند. با توجه به نتایج این معیارها می‌توان در مورد عملکرد الگوریتم‌های ارائه شده بحث کرد. برای توضیح این چهار معیار باید ابتدا اجزای موردنیاز برای محاسبه آنها بررسی شوند. این اجزا شامل چهار مورد مثبت درست[4]، مثبت نادرست[5]، منفی درست[6] و منفی نادرست[7] می‌باشند.

مثبت درست: نمونه‌های بیماری که درست بیمار تشخیص داده شده‌اند.

مثبت نادرست: نمونه‌های سالمی که نادرست بیمار تشخیص داده شده‌اند.

منفی درست: نمونه‌های سالمی که درست سالم تشخیص داده شده‌اند.

منفی نادرست: نمونه‌های بیماری که نادرست سالم تشخیص داده شده‌اند.

نرخ درستی تشخیص تصاویر ورودی در پردازش تصویر، یکی از مسائل مهمی است که در روش‌های مختلف به‌عنوان یک پارامتر اصلی برای تعیین دقت تشخیص و درستی کار موردنظر قرار می‌گیرد. در حالت کلی برای روش‌های یادگیری با نظارت، به‌طور معمول ۸۰ درصد از داده‌ها در یادگیری ماشین به داده آموزش Train و ۲۰ درصد برای داده آزمون تعلق می‌گیرند.

$$precision=TP / (TP + FP) \quad (۱۴)$$

$$Accuracy=Tp+TN/P+N \quad (۱۵)$$

[1] accuracy
[2] precision
[3] recall
[4] True Positive (TP)
[5] False Positive (FP)
[6] True Negative (TN)
[7] False Negative (FN)

www.SID.ir

Archive of SID

$$sensitivity = TP / (TP + FN) \quad (۱۶)$$

$$specificity = TN / (FP + TN) \quad (۱۷)$$

بحث پیرامون نتایج

ماشین بردار پشتیبان به علت توانایی بالایش در طبقه‌بندی و برتری‌اش نسبت به سایر الگوریتم‌های طبقه‌بندی و رگرسیون، بسیار محبوب شده است. در واقع، ماشین بردار پشتیبان برای طبقه‌بندی دوتایی طراحی شده است؛ بنابراین فرایند آن به سمت حل مسائل کلاسه‌بندی موجود بین رفتار عادی و غیرعادی یا مشکوک، بسیار مفید است. ماشین بردار پشتیبان، یک یادگیری ماشینی است که بردارهای یادگیری را در فضاهایی با ابعاد ویژگی‌های بالا ترسیم می‌کند و به هر بردار، طبقه متناظرش را برچسب می‌زند. اساس کار خوشه‌بند ماشین بردار پشتیبان، خوشه‌بندی داده‌ای خطی است و در بخش خطی داده‌ها، مهم‌ترین نکته این است که خط انتخاب شده، بیشترین حاشیه اعتماد را داشته باشد. یافتن بهترین معادله خطی برای داده‌ها از طریق روش‌های QP انجام می‌شود که روش‌های معروفی در حل مسائل دارای محدودیت هستند. قبل از عملیات تقسیم‌بندی خطی برای توانمندسازی ماشین برای طبقه‌بندی اطلاعات پیچیده‌تر، داده‌ها از طریق عملیات Phi به فضایی با بعد بالاتر، منتقل می‌شوند. مشکلات به‌وجودآمده در ابعاد بالاتر را می‌توان از طریق قضیه دوتایی لاگرانگ برای تبدیل مسئله کمینه موردنظر به شکل دوتایی حل کرد. بدین روش به جای استفاده از روش پیچیده‌ای مثل Phi می‌توان از عملیات ساده‌تری مثل کرنل استفاده کرد که بردار مضربی از Phi می‌باشد. فرایند توسعه مدل ماشین بردار پشتیبان، شامل سه مرحله است: پیش‌پردازش داده‌ها، توسعه مدل، استخراج مدل و در نهایت اجرای آن. این الگوریتم قابل مقیاس‌گذاری است و حجم داده‌ها در عملکرد آن تأثیری ندارد؛ بنابراین پیچیدگی طبقه‌بندی، بستگی به ابعاد فضای ویژگی‌ها ندارد؛از این رو در مقایسه با شبکه‌های عصبی، از این روش می‌توان برای یادگیری مجموعه بزرگی از مدل‌ها استفاده کرد. این الگوریتم براساس ساختار به حداقل‌رسانی ریسک، طراحی شده و کاربردهای زیادی دارد و بسیار موفق بوده است. با توجه به موارد گفته‌شده، درخت تصمیم، گزینه مناسبی برای تشخیص هشدارهای نادرست و طبقه‌بندی آنها می‌باشد. فایل Alert دربردارنده دیتا ست مربوط به یک هفته کاری از پایگاه داده DARPA KDD cup 99 می‌باشد. این فایل بعد از اینکه توسط نرم‌افزار اسنورت، ارزیابی شد با استفاده از فایل Tcp list برچسب‌گذاری می‌شود و به‌عنوان فایل ورودی در اختیار وکا قرار می‌گیرد تا الگوریتم درخت روی این فایل اعمال گردد. بعد از اعمال الگوریتم ماشین بردار پشتیبان، نتایجی که در ادامه ارائه شده است به‌دست می‌آید.

در قسمت اجرا شده با نرم‌افزار متلب، از ماشین بردار پشتیبان برای تشخیص نفوذ استفاده شده است که برای این هدف، ابتدا داده، نرمالایز شد و سپس برای استفاده از برنامه متلب، ویرایش شد و استرینگ‌ها به عدد تبدیل شدند. فایل ویرایش‌شده داده‌ها در KDDCup99.txt ارائه شده است. برای استفاده از این فایل، دو فایل با نام‌های traindata.txt و testdata.txt ایجاد شده است که هر دوی این فایل‌ها مجموعه‌ای از داده‌های موجود در KDDCup99.txt است. در اجرای انجام‌شده در ابتدا داده‌های آموزشی ماشین بردار پشتیبان با هدف آموزش روش پیشنهادی و همین‌طور در ادامه برای آزمون صحت روش ماشین بردار پشتیبان موردنظر مورد استفاده قرار می‌گیرند. در نهایت، پس از انجام این مراحل، خروجی به‌دست‌آمده شامل تعداد نفوذها و غیرنفوذها می‌باشد در داده‌های آموزشی با استفاده از یادگیری ماشین بردار پشتیبان می‌باشد.

در جدول ۱ نتایج به‌دست‌آمده از سایر روش‌ها در تشخیص نفوذ که توسط سایر محققان صورت گرفته است در مقایسه با روش ارائه‌شده در این مقاله با استفاده از ماشین بردار پشتیبان ارائه شده است و نتایج، نشان‌دهنده یک بهبود در دقت شناسایی در روش مبتنی بر ماشین بردار پشتیبان نسبت به سایر روش‌ها می‌باشد.

www.SID.ir

Archive of SID

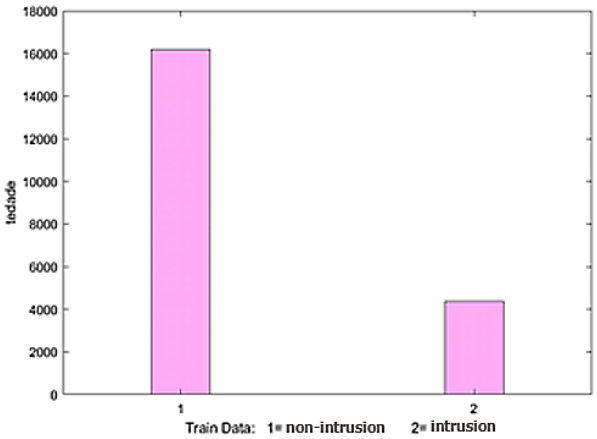
جدول ۱. مقایسه دقت شناسایی روش‌های مختلف

الگوریتم	دقت شناسایی
روش تابع عضویت اصلی[1]	۹۸/۷۸
مدل ترکیبی گاوس کلاس نرمال	۹۷/۲۷
IID	۹۳/۵۵
ماشین بردار پشتیبان	۹۲ .۳۰
xcsسیستم طبقه‌بند	۹۱
مدل ترکیبی گوسی داده‌های نرمال	۸۶/۵۴
درخت NB	۸۲/۰۲
درخت تصادفی	۸۱/۵۹
جنگل تصادفی[2]	۸۰/۶۷
پرسپترون چندلایه[3]	۷۷/۴۱
درخت بیز ساده[4]	۷۶/۵۶

جدول ۲. ارائه پارامترهای الگوریتم ماشین بردار پشتیبان

پارامترها	ماشین بردار پشتیبان
Correctly Classified Instances	٪۹۸/۴۱۷۱
Incorrectly Classified Instances	٪۱/۵۸۲۹
Kappa statistic	٪۰/۹۶۸۱
Mean absolute error	٪۰/۰۲۸۶
Root mean squared error	٪۰/۱۲۱۷
Relative absolute error	٪۵/۷۵۲۸
Root relative squared error	٪۲۴/۴۰۵۵
Total Number of Instances	۱۲۵۹۷۳

در جدول ۲ مقادیر به‌دست‌آمده از اجرای روش پیشنهاد شده در محیط وکا ارائه گردیده است.



شکل ۷. مقایسه تشخیص نفوذ و غیرنفوذ در داده‌های آموزشی

[1] Membership function (MMF)
[2] Random forest
[3] Multilayer perceptron(MLP)
[4] Naive Bayes

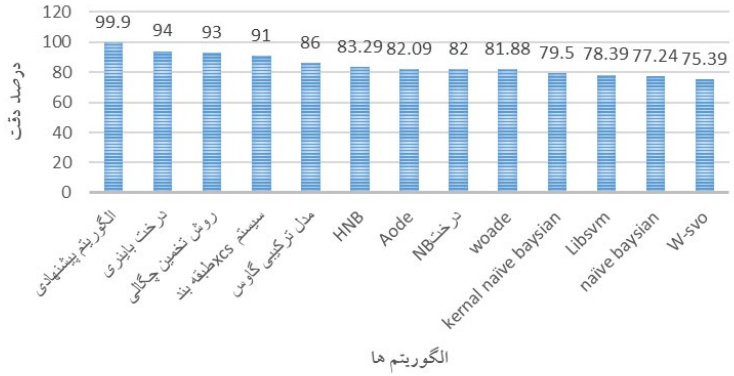
www.SID.ir

Archive of SID

در شکل ۷ نتایج حاصل از میزان تشخیص نفوذ و غیرنفوذ در روش پیشنهاد شده مبتنی بر روش ماشین بردار پشتیبان ارائه شده است در این نمودار میزان نفوذ و غیر نفوذ در داده‌های آموزشی که ۷۰ درصد از کل داده‌های استفاده شده را تشکیل می‌دهند استفاده شده است. هدف از تشکیل این مقادیر و نمودار آموزش ماشین بردار پشتیبان برای تشخیص درست حملات می‌باشد.

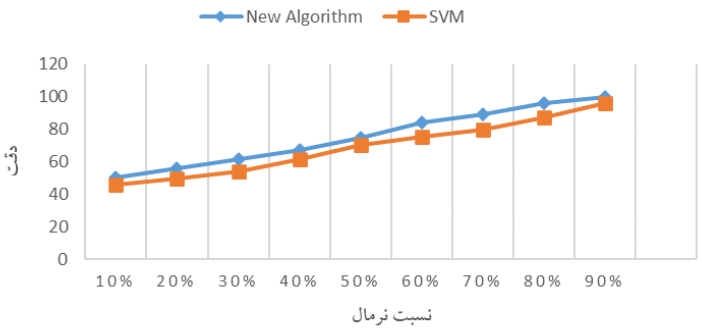
نتایج عددی و بررسی صحت نتایج

در این قسمت، مقادیر عددی به‌دست‌آمده ارائه شده است که نشان‌دهنده میزان افزایش دقت روش پیشنهادی نسبت به روش‌های قبلی است. همین‌طور نمودارهای مقایسه‌ای در ادامه ارائه شده است. در نمودارهای زیر مقادیر نرخ تشخیص و دقت در روش‌های مختلف، مقایسه شده است.



شکل ۸. نمودار مقایسه مقادیر دقت برای الگوریتم‌های مختلف

در شکل ۸ میزان دقت تشخیص نفوذ در سیستم‌های تشخیص نفوذ با استفاده از روش پیشنهادی و سایر روش‌ها که توسط سایر محققان و در مقالات متعدد صورت گرفته است، مقایسه‌ای صورت گرفته است و با توجه به نتایج به‌دست‌آمده روش ماشین بردار پشتیبان که در این مقاله به‌کار رفته به دلیل قدرت بسیار بالای این روش در طبقه‌بندی و دسته‌بندی بسیار خوب عمل کرده و بهبود قابل‌توجهی در مقایسه با روش‌های مشابه ایجاد کرده است.

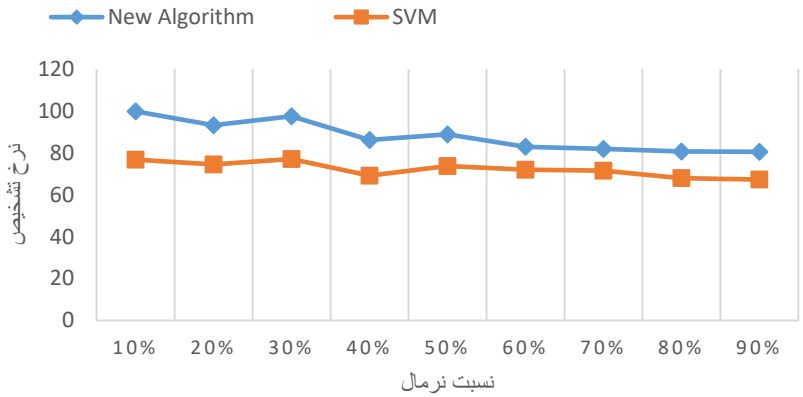


شکل ۹. نمودار خطی مقایسه دقت شناسایی برای نسبت‌های مختلف

www.SID.ir

Archive of SID

نمودار موجود در شکل دقت‌های مختلف به‌دست‌آمده در نسبت‌ها در روش پیشنهاد شده در این مقاله که مبتنی بر روش ماشین بردار پشتیبان است را نشان می‌دهد که طبق نمودار، روند دقت روش پیشنهادی، نسبت به مقادیر نسبی اعلام‌شده، سیر صعودی دارد و از دقت مناسبی در سیستم‌های تشخیص نفوذ برخوردار هستند.



شکل ۱۰. نمودار مقایسه نرخ تشخیص هشدارها برای نسبت‌های مختلف

در شکل ۱۰ مقایسه‌ای در مقادیر به‌دست‌آمده از نرخ تشخیص در روش پیشنهادی در این مقاله که برمبنای روش ترکیبی ماشین بردار پشتیبان و درخت ID3 که برای انتخاب و خوشه‌بندی از این دو روش استفاده می‌کند، با روش ماشین بردار پشتیبان به‌تنهایی صورت گرفته است. طبق نتایج به‌دست‌آمده از نرخ تشخیص، این روش به دلیل ترکیب با درخت و خوشه‌بندی داده‌ها توسط درخت ID3 کارایی بهتری نسبت به روش ماشین بردار پشتیبان ساده پیدا کرده و نتایج قابل‌قبولی را برای سیستم‌های تشخیص نفوذ ارائه می‌دهند.

نتیجه گیری

در این مقاله، سیستم جدید برای انتخاب خصیصه و مدیریت هشدار که قابلیت اجرای فعال را داشت، شبیه‌سازی شد. نتایج به‌دست‌آمده از اجرای آن ارائه شدند. نتایج به‌دست‌آمده از سیستم جدید با دیگر روش‌های مدیریت هشدار مقایسه گردید و مشاهده شد که سیستم جدید دارای دقت، سرعت و کارایی بسیار بالایی نسبت به دیگر روش‌های مدیریت هشدار مبتنی بر خوشه‌بندی است. همچنین با توجه به نتایج به‌دست‌آمده، مشاهده می‌شود که این سیستم قادر به دسته‌بندی فعال هشدارهای سیستم‌های تشخیص نفوذ می‌باشد. دقت اجرای الگوریتم برای به‌دست آوردن هشدارهای درست به ۹۹/۹ درصد افزایش یافته است و نتایج به‌دست‌آمده از مقایسه‌ای که با روش‌های پیشین انجام شده است حاکی از این قضیه است که دقت شناسایی نسبت به روش‌های پیشین، بهبود یافته است. با توجه به این نکته که هر روشی مزایا و معایب خاص خود را دارد روش پیشنهادی نیز در کنار مزایایی که بیان شد معایبی نیز دارد؛ از جمله معایبی که بر این روش وارد می‌باشد به این شرح است که محدودیت‌های ذاتی دارند؛ برای مثال، هنوز مشخص نشده است که به‌ازای یک تابع نگاشت، پارامترها را چگونه باید تعیین کرد، همین‌طور ماشین‌های مبتنی بر بردار پشتیبان به محاسبات پیچیده و زمان‌بر نیاز دارند و به دلیل

www.SID.ir

Archive of SID

پیچیدگی محاسباتی، حافظه زیادی نیز مصرف می‌کنند، از طرفی، داده‌های گسسته و غیرعددی هم با این روش سازگار نیستند و باید تبدیل شوند.

Reference

1. Liao.H, Lin.Ch, Lin.Y , Tung.K, (2013) , Intrusion detection system, A comprehensive review, Journal of Network and Computer Applications, Vol. 36, pp.16-24.
2. Teodoro.P.G , Verdejo.J.D , Fernandez.G.M , Vazquez.E , (2009) , Anomaly-based network intrusion detection, Techniques, systems and challenges, Computers & Security,pp.28-18.
3. Kim.G , Lee.S, Kim.S , (2014) , A novel hybrid intrusion detection method integrating anomaly detection with misuse detection", Expert Systems with Applications, Vol. 41, PP. 1690–1700.
4. Zarkasb.O , Shiri.S, (2009) , detection of penetration in data base using coupling of letters event , fifth international conference on IT management , Tehran .(in Persian)
5. Maiwald.E , (2013) , "Security a Beginners Guide", Mc Graw Hill , New York ,pp:149.
6. Brenton.C , Hunt.C , (2002) , Mastering Network Security", Sybex, California ,ISBN: 0-7821-4142-0, pp:420.
7. Ning.P , Cui.Y , Reeves.D , (2002) , Constructing attack scenarios through correlation of intrusion alerts, Proceedings of the 9th ACM conference on Computer and communications security, ACM, pp: 245-254.
8. Wood.M , Erlinger.M , (2007) , Intrusion Detection Message Exchange Requirements", draft-ietf-idwg-requirements ,pp:8.
9. Sweeney.M , Baumrucker.C , Burton.D , & I. Dubrawsky, (2003) , "Cisco Security Professional´s Guide to Secure Intrusion Detection Systems (IDS)", Syngress, US , ISBN: 9781932266696 , pp:370.
10. Mchugh.J , (2000) , DARPA intrusion detection system evaluations as performed by Lincoln Laboratory , HYPERLINK "https://dl.acm.org/journal/tissec"ACM Transactions on Information and System Security HYPERLINK "https://doi.org/10.1145/382912.382923 " doi:10.1145/382912.382923.
11. Salehpour.N ,Nazari farokhi.M , Nazari farokhi.E , (2013) , presenting a method on vector machine for computer network intrusion detection, AFTA magazine, 2 (6), pp:51 -64.(in Persian)
12. Hashemi.S , (2013) , efficiency of SVM and PCA to enhance intrusion detection systems, Journal of Asian Scientific research, 3(4) , pp:381-395 . (in persian)
13. Gollmann.D , (2013) , Computer Security, Wiley , New Jersey, pp:160. ISBN: 978-1-118-80132-1
14. Hamidi.A , Ziaie.M , (2010) , introduction of penetrate detecting systems Snort , APA professional lab , Ferdowsi university of Mashhad .(in Persian)

www.SID.ir