مجله علمی-پژوهشی
رایانش نرم و فناوری اطلاعات
جلد 1، شماره 3، سال 1391

افزایش نرخ کشف نفوذ به شبکه‌های کامپیوتری با استفاده از درخت‌های تصمیم

پریسا جعفرزاده[1]، شهرام جمالی[2]
[1] کارشناسی‌ارشد، دانشگاه آزاد اسلامی، واحد زنجان،گروه کامپیوتر، زنجان، ایران، jafarzadeh_parisa@yahoo.com
[2] دانشیار دانشگاه محقق اردبیلی، گروه مهندسی کامپیوتر، اردبیل، ایران، jamali@uma.ac.ir

چکیده

امنیت، یکی از مسائل مهم در سیستم‌های رایانه‌ای مدرن امروزی است؛ از جمله چالش‌های مهم در این سیستم‌ها، تشخیص نفوذ است؛ نفوذ، به عنوان مجموعه‌ای از فعالیت‌ها تعریف می‌شود که هدف این فعالیت‌ها، به خطر انداختن یکپارچگی، قابلیت اطمینان سیستم و دسترسی غیرمجاز به منبع خاص می‌باشد. لذا سیستم‌های تشخیص نفوذ یکی از تکنیک‌هایی است که برای حفظ امنیت در شبکه‌های رایانه‌ای استفاده می‌شود. یکی از مشکلات این سیستم‌های امنیتی،گزارش نادرست هشدار نفوذ به سیستم است؛ این مقاله با استفاده از درخت تصمیم یک مکانیزم کشف نفوذ به شبکه طراحی می‌کند. نتایج شبیه‌سازی نشان می‌دهد که نه تنها در این روش مدت زمان فاز آموش کاهش می یابد، بلکه نرخ هشدار نادرست و نرخ تشخیص نفوذ نیز بهبود نسبی پیدا می‌کند.

***کلید واژه‌ها**: داده‌کاوی، درخت تصمیم، سیستم تشخیص نفوذ.*

1. مقدمه

رشد روز افزون استفاده از خدمات شبکه‌های رایانه‌ای از یک سو و حمله به این شبکه‌ها از سوی دیگر باعث شده است که تشخیص نفوذ به عنوان یک زمینه‌ی تحقیقاتی مهم در مسأله‌ی امنیت این شبکه‌ها مطرح شود.

سیستم تشخیص نفوذ[1]، ابزار امنیتی مؤثر و کارآمدی است که در شبکه‌های رایانه‌ای قرار می‌گیرد و با استفاده از یک‌سری قوانین از پیش تعریف‌شده، دسترسی کاربران را بررسی و محدود می‌نماید. این قوانین بر پایه‌ی دانش افراد متخصص حاصل شده است.

سیستم‌های تشخیص نفوذ، تمام نفوذهای کاربران را مورد شناسایی قرار می‌دهد و فعالیت‌های لازم را برای توقف حملات اتخاذ می‌کند. مسأله‌ی تشخیص نفوذ به طور وسیع در زمینه‌ی امنیت سیستم‌های رایانه‌ای مورد مطالعه قرار گرفته است و اخیراً در زمینه‌های یادگیری ماشین و داده‌کاوی توجه زیادی به آن شده است. سیستم‌های تشخیص نفوذ، حملات شناخته‌شده و حملات بالقوه را در ترافیک شبکه و یا داده‌های بازرسی[2] و یا فایل‌های گزارش (ثبت وقایع[3]) ثبت‌شده به وسیله‌ی میزبان‌ها جستجو می‌کند و اقدامات لازم را جهت حفاظت از سیستم انجام می‌دهد. لازم به ذکر است که اولین منبع اطلاعات که سیستم‌های تشخیص نفوذ از آنها استفاده می‌کنند داده‌های بازرسی سیستم‌عامل می‌باشند. داده‌های بازرسی توسط بخشی به نام بخش بازرسی که

زیرمجموعه‌ی سیستم عامل است تهیه می‌شود. داده‌های بازرسی دربرگیرنده‌ی اطلاعاتی درباره‌ی فعالیت‌های سیستم می‌باشد. این اطلاعات برحسب زمان مرتب شده‌اند و دریک یا چند فایل به نام فایل بازرسی ذخیره می‌شوند. هرفایل بازرسی حاوی مجموعه‌ای از رکوردهای بازرسی است که هر یک بیانگر یک رویداد در سیستم هستند. این رکوردها توسط فعالیت‌های کاربر یا فرآیندها ایجاد می‌شوند.

اولین نیاز سیستم‌های تشخیص نفوذ، وجود منبع اطلاعات است. این منبع اطلاعات، به عنوان تولیدکننده‌ی رویداد در نظرگرفته می‌شود. درسیستم‌های تشخیص نفوذ، این منابع اطلاعاتی با توجه به مکانشان دسته‌بندی می‌شود. با توجه به این معیار، دو دسته‌بندی کلی برای سیستم‌های تشخیص نفوذ وجود دارد[1]:

مبتنی برشبکه[4]: در این دسته، بسته‌های عبوری درسطح شبکه به عنوان منبع اطلاعات جمع‌آوری می‌شوند؛ این عمل، با قرار دادن کارت شبکه در حالت بی‌قاعده[5] صورت می‌گیرد. در این حالت چون سرآیند[6] بسته‌های شبکه مورد بررسی قرار می‌گیرد، محدودیت نگاه به سرآیند بسته‌ها باعث می‌شود این سیستم‌ها قابلیت تشخیص دامنه‌ی محدودتری ازحملات را داشته باشند؛ لازم به ذکر است که منظور از حالت بی‌قاعده این است که در حالت طبیعی، یک سیستم بر روی شبکه، تنها ترافیکی که به طور مستقیم به آن ارسال می‌شود را می‌خواند و پاسخ می‌دهد، ولی در حالت بی‌قاعده، سیستم، تمام ترافیک شبکه را می‌خواند.

مبتنی بر میزبان[7]: داده‌های این نوع سیستم‌ها از فعالیت‌های متنوع کامپیوترهای میزبان که شامل رکوردهای بازرسی سیستم‌عاملی، فایل‌های ثبت رخداد سیستمی، اطلاعات برنامه‌های کاربردی و ... هستند، فراهم می‌شود. اطلاعات از منابع موجود در این میزبان‌ها جمع‌آوری می‌شود. همان‌گونه که اشاره شد، این منابع شامل فایل‌های ثبت وقایع و داده‌های بازرسی هستند.

از آنجا که این سیستم‌ها از منابع خود دستگاه استفاده می‌کنند، باعث ایجاد سربار اضافی بر روی دستگاه می‌شوند، ولی از نظر تئوری، قابلیت تشخیص تقریباً تمام حملات را دارند. این سیستم‌های تشخیص نفوذ می‌توانند در مورد نفوذی بودن یا نبودن یک کامپیوتر میزبان به دقت داوری کنند، زیرا داده‌های آنها از داده‌های مورد بازرسی سیستم و فایل‌های ثبت رخدادها در سیستم‌عامل به دست می‌آیند و در مقایسه با سیستم‌های تشخیص نفوذ مبتنی بر شبکه می‌توانند حملات داخل شبکه و میزبان‌های نفوذی را شناسایی نمایند.

بعد از معرفی منابع اطلاعاتی، باید نوع تحلیل‌گر شبکه مشخص شود. در تحلیل‌گر، اطلاعات از منابع اطلاعاتی استخراج می‌شوند و با توجه به سیاست‌های امنیتی، انواع حملات مورد بررسی قرار می‌گیرند. بر اساس تحلیل‌گر، دو نوع دسته‌بندی داریم:

1) تشخیص رفتارغیرعادی: در این گونه سیستم‌ها تشخیص حمله به این صورت است که رفتار عادی سیستم را مدل می‌کنیم و رفتار و قواعد خاصی برای آنها در نظرگرفته می‌شود؛ رفتارهایی که از این الگوها پیروی کنند، عادی می‌باشد و رفتارهایی که مغایر با این الگوها هستند، به عنوان رفتار غیرعادی در نظر گرفته می‌شوند. این سیستم‌ها معمولاً دارای تعداد هشدار نادرست بالا هستند، ولی قابلیت تشخیص حمله‌های ناشناخته را نیزبه سیستم می‌دهند. در این گونه دسته‌بندی، سیستم یک مرز نرمال از الگوهای متعارف استفاده از سیستم برای خود ترسیم و تعریف می‌کند. هر رفتار و رویدادی که به میزان زیادی از این الگوها دور باشد به عنوان اخلالگر محتمل بر شبکه در نظر گرفته می‌شود. آنچه خلاف قاعده و نابهنجار فرض می‌شود می‌تواند متغیر باشد. اما معمولاً رویدادی که با تناوبی بیشتر یا کمتر از دو مرتبه انحراف از آمار نرمال به وقوع بپیوندد نابهنجار فرض می‌شود. چند مثال برای این شرایط عبارت است از:

کاربری که به جای یک یا دو بار ورود و خروج از سیستم در طول روز 20 بار این کار را انجام داده است. کامپیوتری که در ساعت 2 بعد از نیمه شب مورد استفاده قرار گرفته است، در صورتی که قرار نبوده بعد از ساعت اداری روشن باشد. در یک سطح دیگر این تکنیک می‌تواند الگوهایی در مورد کاربران از جمله برنامه‌هایی که به اجرا در می‌آورند را مورد استفاده قرار دهد. مثلاً اگر کاربری از بخش گرافیک یک سازمان ناگهان شروع به دست‌یابی به برنامه‌های حسابداری یا کامپایل کردن کد بنماید، سیستم می‌تواند یک هشدار به مدیر مسؤول امنیتی[8] شبکه ارسال کند[2].

2) تشخیص سوء استفاده[9]: در این حالت، الگوهای نفوذ از پیش ساخته‌شده، به عنوان قانون نگهداری می‌شوند. در واقع، در این حالت، هدف، تشخیص حملات شناخته‌شده به همراه تغییرات کوچک در آنها می‌باشد. از مزایای این روش، دقت در تشخیص حملات شناخته‌شده وکاهش هشدارهای نادرست می‌باشد. این سیستم، توانایی تشخیص حملات ناشناخته را ندارد[2].

سیستم‌های تشخیص نفوذ، از روش‌های مختلفی همانند درخت‌های تصمیم‌گیری[10]، الگوریتم ژنتیک[11]، منطق فازی[12]، خوشه‌بندی[13]، شبکه‌ی بیزین[14] و ... برای کشف نفوذ استفاده می‌کند.

مابقی این مقاله به این ترتیب سازماندهی شده است: در بخش 2، به طور مختصر به مفاهیم داده‌کاوی و چرایی استفاده از آن می‌پردازیم. در بخش 3، کاربرد داده‌کاوی را در ساخت سیستم‌های تشخیص نفوذ کارآمد بیان می‌کنیم. در ادامه و در بخش 4، به راهکار یادگیری ماشین که ارتباط نزدیکی با تکنیک‌های داده‌کاوی دارد، اشاره می‌کنیم. همچنین، به الگوریتم C4.5 می‌پردازیم که یکی از روش‌های یادگیری ماشین است. در بخش 5، به شبیه‌سازی و ارزیابی نتایج حاصل می‌پردازیم. در نهایت، در بخش 6، نتیجه‌گیری خود را از این مقاله ارائه می‌کنیم.

2. داده‌کاوی

استخراج دانش از اطلاعات می‌باشد. در اینجا منظور از استخراج اطلاعات، دستیابی به اطلاعاتی است که قبلاً بدیهی نبوده و برای ما ناشناخته بوده‌اند. قبل از آغاز اجرای عملیات داده‌کاوی بر روی اطلاعات، لازم است که پیش‌پردازش‌هایی روی داده‌های خام و انبوه انجام دهیم؛ این عملیات پیش‌پردازش، شامل انتخاب خصیصه‌های مفید از داده‌ها، کاهش حجم و بُعد داده‌ها به منظور کاهش زمان محاسبات داده‌کاوی، نرمال‌سازی داده‌ها، گسسته‌سازی داده‌ها و ... می‌باشد[3].

بعد از انجام کاوش و استخراج الگوهای مفید از داده‌ها، ممکن است نیاز به انجام پردازش‌هایی روی این الگوها باشد. مجموعه‌ی این اعمال را استخراج دانش از پایگاه‌های داده نامیده می‌شود.

همان‌گونه که در شکل 1 مشاهده می‌شود، فرآیند کشف دانش متشکل از چند مرحله می‌باشد که از داده‌های خام، اطلاعات مفید را به دست می‌دهد. این مراحل عبارتند از:

پاکسازی داده‌ها[15]: این مرحله با عنوان پیرایش داده‌ها نیز شناخته می‌شود. در این فاز، داده‌های اضافی و نامربوط از مجموعه‌ی داده‌های موجود حذف می‌شوند.

تمامیت داده‌ها[16]: دراین مرحله، تصمیم‌گیری بر روی داده‌ها متناسب با تحلیل، انجام می‌شود و این داده‌ها از مجموعه داده‌ها استخراج می‌شوند.

تبدیل صورت داده‌ها[17]: این مرحله با عنوان تثبیت داده نیز شناخته می‌شود که در این فاز، داده‌های انتخاب‌شده به صورتی متناسب جهت فرآیند داده‌کاوی تبدیل می‌شوند.

داده‌کاوی[18]: این مرحله بسیار مشکل می‌باشد و در آن از تکنیک‌های هوشمند برای استخراج الگوهای مفید بالقوه، استفاده می‌شود.

ارزیابی ارزش الگوها[19] : در این مرحله الگوهای صریح به منظور کشف دانش، شناسایی می‌شوند.

اکتشاف دانش[20]: در این مرحله‌ی نهایی، دانش مورد نظر کشف‌شده، به صورت بصری به کاربر ارائه می‌شود. این مرحله‌ی مهم از تکنیک‌های مصوّرسازی برای کمک به کاربران در درک و تفسیر نتایج داده‌کاوی، استفاده می‌کند.

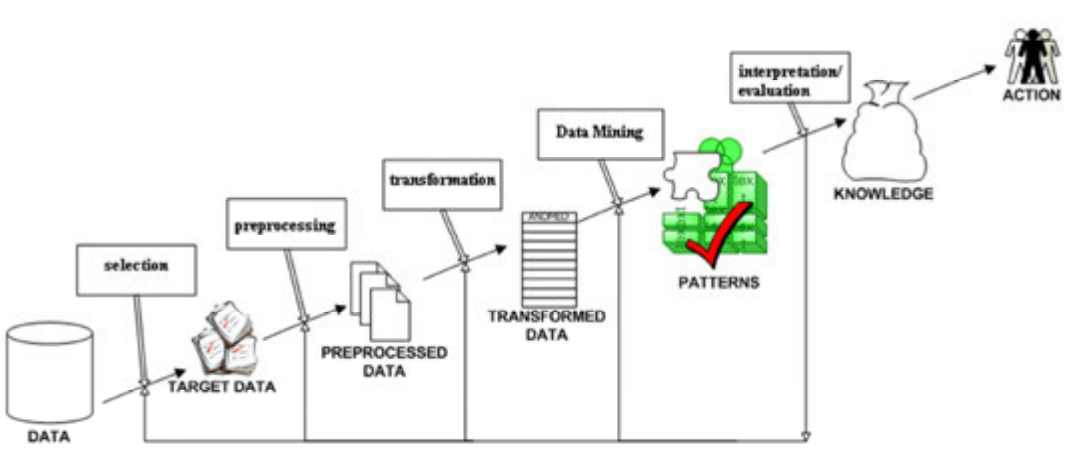
فرآیند اکتشاف علوم، یک فرآیند تکرارشونده است. زمانی که علوم کشف و به کاربران ارائه شد، سنجش‌هایی جهت ارزشیابی انجام می‌شود و امکان پاک شدن مجدد علوم به دست آمده در طی عملیات داده‌کاوی با داده‌های جدید و یا ترکیبی از منابع داده‌ای جدید وجود دارد و ارائه‌ی نتایج بهتری را به ارمغان می‌آورد.

3. کاربرد داده‌کاوی در تشخیص نفوذ

نفوذ، در حقیقت به مجموعه اقدامات غیرقانونی اطلاق می‌شود که صحت، محرمانگی و یا دسترسی به یک منبع را به خطر می‌اندازد. عوامل نفوذی را می‌توان به دو گروه خارجی و داخلی تقسیم کرد نفوذی‌های خارجی کسانی هستند که اجازه‌ی استفاده از منابع سیستم را ندارند، اما سعی می‌کنند آنها را مورد دستیابی قرار دهند و نفوذی‌های داخلی کسانی هستند که برای دستیابی به سیستم اختیارات محدودی دارند، اما سعی می‌کنند به منابعی که اجازه‌ی استفاده از آنها را ندارند، دسترسی پیدا کنند. به منظور مقابله با نفوذکنندگان به سیستم‌ها و شبکه‌های کامپیوتری توسط هر دو دسته کاربران داخلی و حمله‌کنندگان خارجی، روش‌های متعددی پیشنهاد شده است که تکنیک‌های تشخیص نفوذ نامیده می‌شوند. هدف از تشخیص نفوذ این است که استفاده‌ی غیرمجاز، سوءاستفاده و آسیب رساندن به سیستم‌ها و شبکه‌های کامپیوتری شناسایی و با آنها مقابله شود. یکی از پرکاربردترین تکنیک‌های تشخیص نفوذ، یادگیری ماشین می‌باشد که در بخش 4 به طور مختصر معرفی می‌شود.

4. یادگیری ماشین

یادگیری ماشین، شاخه‌ای از هوش مصنوعی است که در حقیقت برخی از تکنولوژی‌هایی را که در آنها کامپیوتر، شرایط یادگیری اتوماتیک را دارد، گسترش می‌دهد. بحث اصلی در یاگیری ماشین این است که چگونه از کامپیوترها و روش‌های آماری استفاده کنیم تا بتوانیم اطلاعات مفید را از میان حجم انبوه داده‌ها استخراج کنیم. بنابراین یادگیری ماشین و داده‌کاوی هر دو با روش‌های آماری و تئوری دانش کامپیوتر در ارتباط هستند. تکنولوژی‌های رایج یادگیری ماشین شامل موارد زیر هستند [5]: تئوری تصمیم بیز[21]، روش‌های چند متغیری[22]، خوشه‌بندی، درخت‌های دسته‌بندی[23]، تفکیک خطی[24]، مشاهدات چندلایه[25]، مدل‌های محلی[26]، مدل‌های مارکوف پنهان[27]، یادگیری تقویت[28]. تکنیک‌های یادگیری ماشین در شرایطی مناسب است که هیچ گونه دانش اولیه در مورد الگوهای داده‌ها وجود ندارد؛



شکل (1): مراحل داده‌کاوی و استخراج دانش [4]

به همین دلیل گاهی به این روش‌ها، پائین به بالا می‌گویند. مزیت مهم این تکنیک‌ها این است که معمولاً به انسان‌های خبره برای تعیین ملزومات مورد نظر به منظور تشخیص نفوذ نیازی نیست به همین دلیل بسیار سریع عمل می‌کنند و مقرون به صرفه هستند.

در بخش بعد به بررسی درخت تصمیم می‌پردازیم و نوآوری خود را که در مورد انتخاب ویژگی‌های تأثیرگذار مجموعه داده‌ی به کار رفته در فاز آموزش است تشریح می‌کنیم.

5. درخت دسته‌بندی

دسته‌بندی، یک مدل پیشگویی در یادگیری ماشین است و درخت تصمیم نیز نامیده می‌شود که یک گراف با الگوی درختی و شبیه به ساختمان داده‌ی نمودار گردشی[29] است.

درخت تصمیم درختی است که در آن نمونه‌ها را به نحوی دسته‌بندی می‌کند که از ریشه به سمت پائین رشد می‌کنند و در نهایت به گره‌های برگ می‌رسد: هر گره‌ی داخلی یا غیربرگ بایک ویژگی مشخص می‌شود.این ویژگی سؤالی را در رابطه با مثال ورودی مطرح می‌کند.

درهر گره‌ی داخلی به تعداد جواب‌های ممکن با این سؤال، شاخه وجود دارد که هر شاخه با مقدار آن جواب مشخص می‌شود.برگ‌های این درخت با یک کلاس و یا یک دسته از جواب‌ها مشخص می‌شوند. مهمترین و اساسی‌ترین الگوریتم‌هایی که در درخت‌های تصمیم استفاده می‌شوند ID3 و C4.5 که هر دو متعلق به ساختمان داده‌ی درخت هستند و الگوریتم آنها به صورت زیر می‌باشد[6]:

تمام نمونه‌های داده‌های آموزشی در مکان ریشه‌ی درخت قرار می‌گیرند.

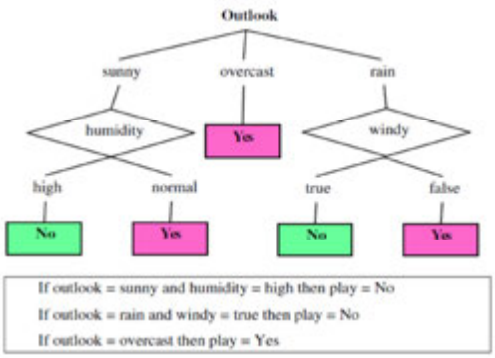
اگر گرهی دارای هیچ داده‌ای نباشد و یا داده‌های یک گره از نوع یکسان باشند، در این صورت گره، یا به عنوان یک گره‌ی برگ خالی می‌باشد و یا به عنوان برگی از نمونه‌های یکسان در نظر گرفته می‌شود. اگر گره دارای نمونه‌های بیشتر از یک نوع باشد، باید بتوانیم طبق یک تابع ارزیابی معین به تمام خصیصه‌های داده‌ها دست یابیم و بتوانیم خصیصه‌های مناسب را انتخاب کنیم. بر مبنای تعداد این خصیصه‌ها، نمونه‌های یک گره به N قسمت تقسیم می‌شوند و هر قسمت به عنوان یک گره‌ی جدید است که به گره‌ی ریشه متصل شده است. این فرآیند، تقسیم گره[30] نامیده می‌شود.

بعد از تقسیم گره، تصمیم‌گیری در مورد برگ بودن گره‌های جدید انجام می‌گیرد. اگر برگ نباشند، این گره‌ها به عنوان ریشه برای زیردرخت‌های جدید

قرار می‌گیرند و برای تشکیل زیردرخت‌های جدید استفاده می‌شوند.

مراحل فوق به روش بازگشتی بسط داده می‌شوند، تا زمانی که تمام گره‌های جدید به عنوان گره‌های برگ باشند. تابع ارزیابی نیز معمولاً شامل تابع آنتروپی[31]، بهره‌ی اطلاعاتی[32]، تابع جینی[33]، پیش‌هرس درخت[34] و پس‌هرس درخت[35] می‌باشد. بعد از محاسبه‌ی مقادیر این توابع، ویژگی با بیشترین مقدار انتخاب می‌شود.

شکل 2، ساختار یک درخت تصمیم‌گیری و قوانین استخراج شده از آن را نشان می‌دهد. خوانندگان این مقاله برای دریافت اطلاعات بیشتر می‌توانند به مرجع [6] مراجعه نمایند.



شکل 2: ساختار یک درخت تصمیم‌گیری و قوانین استخراج‌شده از آن [6]

از مزایای درخت‌های دسته‌بندی می‌توان به موارد زیر اشاره کرد:

1) کارکردن با داده‌های بزرگ و پیچیده: درخت تصمیم در عین سادگی می‌تواند با داده‌های پیچیده به راحتی کار کند و از روی آنها تصمیم بسازد.
2) استفاده‌ی مجدد آسان: در صورتی که درخت تصمیم برای یک مسأله ساخته شد، نمونه‌های مختلف از آن مسأله را می‌توان با آن درخت تصمیم محاسبه کرد.
3) قابلیت ترکیب با روش‌های دیگر: نتیجه درخت تصمیم را می‌توان با تکنیک‌های تصمیم‌سازی دیگر ترکیب کرده و نتایج بهتری به دست آورد.

با توجه به مزایای فوق و بررسی کارهای انجام‌شده [2] در رابطه با تشخیص نفوذ متوجه شدیم که درخت تصمیم دارای نرخ هشدار نادرست کم و نرخ تشخیص نفوذ بالایی است؛ ولی ایراد این روش این است که مدت زمان آموزش این الگوریتم در فاز آموزش نسبتاً زیاد است بنابراین بایستی حجم مجموعه داده‌ای که در فاز آموزش برای ساخت مدل استفاده می‌شود کاهش یابد. به این منظور سعی کردیم که از بین تمام ویژگی‌های موجود در مجموعه داده‌ی به کار رفته در فاز آموزش درخت تصمیم، تنها مهمترین و تأثیرگذارترین ویژگی‌ها را انتخاب کنیم و در ساخت مدل به کار بریم.

همان‌گونه که قبلاً نیز عنوان شد برای کاهش مدت زمان ساخت مدل در این الگوریتم و همچنین افزایش سرعت و جامعیت سیستم مورد نظر، به جای استفاده از تمام 41 ویژگی موجود در مجموعه داده‌ی KDDCup99، از تعدادی ویژگی موجود در این مجموعه داده استفاده کردیم که این ویژگی‌ها با استفاده از الگوریتم‌های انتخاب ویژگی موجود در نرم‌افزار داده‌کاوی Weka انتخاب شدند. این تعداد ویژگی قابل تغییر می‌باشد و بدیهی است که انتخاب ویژگی‌های مناسب، تأثیر به سزایی بر عملکرد خوب سیستم تشخیص نفوذ دارد. نحوه‌ی عملکرد هر کدام از این الگوریتم‌ها به طور مختصر شرح داده می‌شود و برای توضیحات بیشتر به سایت [7Weka] مراجعه شود:

- CfsSubsetEval: دراین روش، ارزش پارامترها به طور مکاشفه‌ای با استفاده از همبستگی ارزیابی می‌شود. در این روش بهترین زیرمجموعه‌ی ویژگی، ویژگی‌هایی هستندکه بیشترین همبستگی را برای پیش‌بینی ویژگی هدف (کلاس) دارند. اما ازسوی دیگر، درجه‌ی

همبستگی بین خود ویژگی‌های موجود در زیر مجموعه کم می‌باشدکه از فرمول رابطه‌ی 1 برای به دست آوردن درجه‌ی شایستگی این زیرمجموعه‌ها استفاده می‌شود[8]:

$$\text{Merit} = \frac{\text{درجه‌ی همبستگی ویژگی مورد نظر با ویژگی هدف}}{\text{درجه‌ی همبستگی ویژگی مورد نظر با ویژگی‌های موجود در زیرمجموعه}} \quad (1)$$

- ChiSquaredAttributeEval: این الگوریتم برای رتبه‌بندی ویژگی‌ها از معیار Chi-Squared استفاده می‌کند. ویژگی‌هایی که رتبه‌ی بالاتری دارند، از قدرت جداکنندگی بیشتری نیز برخوردار هستند.
- ClassifierSubsetEval: این الگوریتم ارزش ویژگی‌ها را با استفاده از کلاس‌بندی که کاربر مشخص می‌کند، ارزیابی می‌کند. به عنوان مثال، کاربر کلاس‌بند OneR را انتخاب می‌کند.
- InfoGainAttributeEval: این روش، یکی از پرکاربردترین روش‌های استاندارد در انتخاب ویژگی‌های مرتبط با ویژگی هدف می‌باشدکه میزان شایستگی ارتباطی که ویژگی با ویژگی هدف را با اندازه‌گیری معیار بهره‌ی اطلاعات[36] مشخص می‌کند. هرچه میزان بهره‌ی اطلاعات یک ویژگی با ویژگی هدف نسبت به بقیه بیشتر باشد، ارتباط ویژگی مورد نظر نیز با ویژگی هدف نسبت به بقیه بیشتر می‌باشد. به بیان دیگر، ویژگی هدف را به وسیله‌ی ویژگی انتخاب شده توسط این مدل با میزان خطای کمتری می‌توان پیش‌بینی کرد. این الگوریتم قابلیت کار با داده‌های عددی و اسمی رادارد. در برخورد با داده‌های عددی پیوسته، ابتدا داده‌های مورد نظر را به صورت باینری در نظر می‌گیرد و سپس معیار بهره‌ی اطلاعات را محاسبه می‌کند. ازتوانایی‌های دیگر این الگوریتم، برخورد با ارزش‌های مفقود شده به عنوان یک ارزش جداگانه می‌باشد. در این روش، ارزش مفقودشده با توجه به نسبت فراوانی ارزش‌های دیگر محاسبه می‌گردد.

بهره‌ی اطلاعات، قدرت دسته‌بندی یک ویژگی را، درمجموعه‌ی داده‌ای بیان می‌کند. هدفاز دسته‌بندی آن است که مجموعه‌ی ورودی به دسته‌هایی تقسیم گردد که هر دسته تنها معرف یک ارزش باشد. به عبارت دیگر خلوص هر دسته بیشینه باشد. ازآنجا که ویژگی‌های دارای بهره‌ی اطلاعات بیشتر، مجموعه داده‌ای را به تعداد قسمت‌های یک متر با حداکثر خلوص تقسیم می‌کنند، بنابراین این ویژگی‌ها برای دسته‌بندی، ویژگی‌های مناسبی به شمار می‌روند.

6. شبیه‌سازی و ارزیابی نتایج

برای شبیه‌سازی، از نرم‌افزار داده‌کاوی Weka استفاده کردیم. این نرم‌افزار، یک واسط همگون برای بسیاری از الگوریتم‌های یادگیری متفاوت، فراهم کرده است که از طریق آن روش‌های پیش‌پردازش، پس‌پردازش و ارزیابی نتایج طرح‌های یادگیری روی همه‌ی مجموعه داده‌های موجود، قابل اعمال است. این سیستم به زبان جاوا نوشته شده است و تقریباً روی هر پلتفرمی اجرا می‌شود و پلت‌فرم جاوایی Weka باعث می‌شود تا اجرای آن روی بیشتر سیستم عامل‌ها امکان‌پذیر باشد. بستر آزمایشی که شبیه‌سازی را روی آن انجام دادیم یک رایانه با سیستم‌عامل لینوکس و نسخه‌ی فدورا 11 بود. این رایانه دارای حافظه‌ی 2 گیگاهرتزی و پردازنده‌ی دو هسته‌ای شرکت اینتل می‌باشد. به منظور شبیه‌سازی نیاز به دو مجموعه داده داریم. یکی از این مجموعه داده‌ها را برای آموزش دادن الگوریتم‌ها استفاده می‌کنیم و مجموعه داده‌ی دیگر را برای آزمایش کردن الگوریتم‌ها استفاده می‌کنیم.

6.1. آماده‌سازی مجموعه داده‌های آموزشی و آزمایشی

مجموعه داده‌های فراوانی به منظور ارزیابی سیستم‌های تشخیص نفوذ وجود دارند. مجموعه داده‌های [9KDD Cup 1999] یکی از پرکاربردترین مجموعه داده‌ها برای ارزیابی سیستم‌ها و مدل‌های ارائه به منظور تشخیص نفوذ می‌باشد. مجموعه داده‌های آموزشی KDDCup1999، شامل 4,898,431 رکورد یا بردار اتصالی است که هر کدام از آنها دارای 41 ویژگی هستند.

این ویژگی‌ها به چهار دسته تقسیم می‌شوند: ویژگی‌های پایه[37]، ویژگی‌های محتوایی[38]، ویژگی‌های ترافیک زمانی[39]، و ویژگی‌های ترافیک میزبان[40]. اتصالات به دو نوع مخرب و عادی تقسیم می‌شوند. مجموعه داده‌های آزمایشی نیز شامل 311,027 رکورد می‌باشد و مجموعه داده‌های آموزشی در کل دارای 23 نوع حمله و مجموعه داده‌های تست نیز شامل 37 نوع حمله هستندکه 14 نوع حمله بیشتر از داده‌های آموزشی هستند. نکته مهم این است که داده‌های آزمایشی دارای احتمال توزیع یکسان با داده‌های آموزشی نیستند و شامل حملات خاصی هستند که در داده‌های آموزشی موجود نیستند، و این امر کار را بسیار واقعی جلوه می‌دهد.

به دلیل تعداد بالای نمونه‌های موجود در این مجموعه داده، نرم‌افزار Weka قادر به نمونه‌برداری از این تعداد رکوردها نمی‌باشد؛ به همین جهت، از نرم‌افزار SQL Server 2010 برای نمونه‌گیری تصادفی استفاده می‌کنیم. چون دارای قدرت پردازشی بالایی می‌باشد. لازم به ذکر است که چون مجموعه داده‌های مورد استفاده در فاز آموزش و آزمایشِ الگوریتم‌ها تا حدودی حجیم هستند، مقدار حافظه‌ی تخصیص‌یافته به پلتفرم جاوا را، که نرم‌افزار Weka روی آن اجرا می‌شود، از 128 مگا بایت به 2 گیگا بایت افزایش دادیم. با اِعمال این تغییر، نرم‌افزار Weka به راحتی می‌تواند با مجموعه داده‌های آموزشی و آزمایشی که برای شبیه‌سازی در نظر گرفته‌ایم، کار کند.

در ادامه با استفاده از تکنیک‌های انتخاب ویژگی، از بین 41 ویژگی، 7 ویژگی را انتخاب کردیم و مجموعه داده‌های مورد استفاده در فاز آموزش و آزمایش را بر اساس این ویژگی‌ها آماده نمودیم. این ویژگی‌ها عبارتست از:

Duration, flag, src_bytes, dst_bytes, num failed_login , is_hot_login و is_guest_login

جدول 1: تعداد رکوردهای کلاس حمله و نرمال در مجموعه داده‌ها

	مجموعه داده‌ی آموزشی	مجموعه داده‌ی آزمایشی
تعداد رکوردهای کلاس حمله	25995	25006
تعداد رکوردهای کلاس عادی	81504	6097
تعداد کل نمونه‌ها	107499	31103

همان‌گونه که در جدول 1 مشاهده می‌کنیم، مشخصات مجموعه داده‌های مورد استفاده برای شبیه‌سازی نشان داده شده است.

6.2. معیارهای ارزیابی

کارآیی سیستم‌های تشخیص نفوذ و حتی در حالت کلی‌تر کارایی دسته‌بندی‌کننده‌ها، غالباً از طریق دو معیار نرخ تشخیص و نرخ خطای مثبت نادرست ارزیابی می‌شود. برای ارائه تعریف این دو معیار ابتدا باید به تعریف دو نوع خطای ممکن در سیستم‌های تشخیص نفوذ بپردازیم. این دو نوع عبارتند از:

- خطای مثبت نادرست: اگر یک اتصال نرمال به عنوان حمله شناسایی شود، یک خصای مثبت نادرست رخ داده است.
- خطای منفی نادرست: اگر یک اتصال با برچسب حمله به عنوان یک اتصال نرمال شناسایی شود، یک خطای منفی نادرست رخ داده است.

با توجه به تعاریف فوق داریم:

$$\text{False positive rate} = \frac{\text{Number of false positives}}{\text{Total number of attack connections}} \quad (2)$$

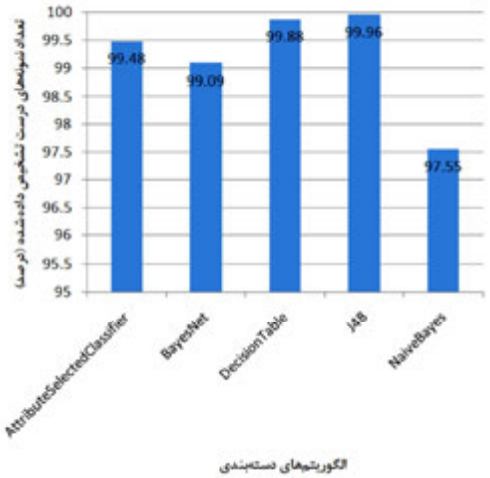
$$\text{False negative rate} = \frac{\text{Number of false negatives}}{\text{Total number of attack connections}} \quad (3)$$

$$\text{Detection rate} = 1 - \frac{\text{Number of false positives}}{\text{Total number of attack connections}} \quad (4)$$

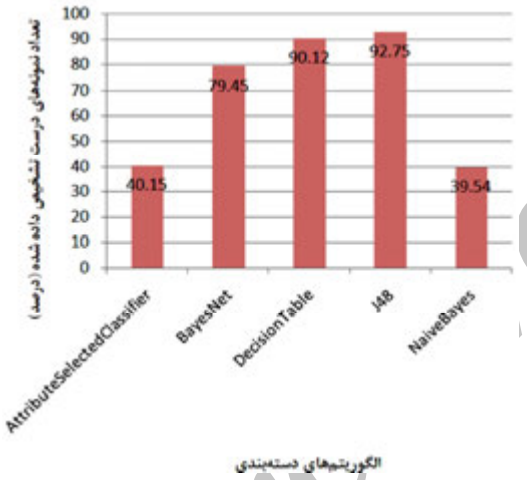
بنابراین نرخ خطای نادرست مثبت (منفی) عبارت است از درصدی از اتصالات نرمال (حمله) که به غلط، برچسب حمله (نرمال) به آنها نسبت داده می‌شود. در آزمایشات انجام شده، از دو معیار نرخ تشخیص و نرخ خطای مثبت نادرست برای ارزیابی راهکار ارائه شده استفاده شده است.

6.3. ارزیابی تشخیص نفوذ مبتنی بر درخت تصمیم

همان‌گونه که در شکل (3) مشاهده می‌شود، الگوریتم C4.5 (J48) نسبت به سایر الگوریتم‌های دسته‌بندی، دارای نرخ تشخیص حملات کمی بیشتر در زمان آموزش می‌باشد. این الگوریتم کارآیی خود را در زمان آزمایش نشان می‌دهد؛ و همان طور که در شکل 4 واضح است، الگوریتم C4.5 بیشترین دقت را در تشخیص حملات از خود نشان می‌دهد.



شکل (3): درصد دقت نمونه‌هایی که در فاز آموزش الگوریتم‌ها به درستی تشخیص داده شده‌اند.

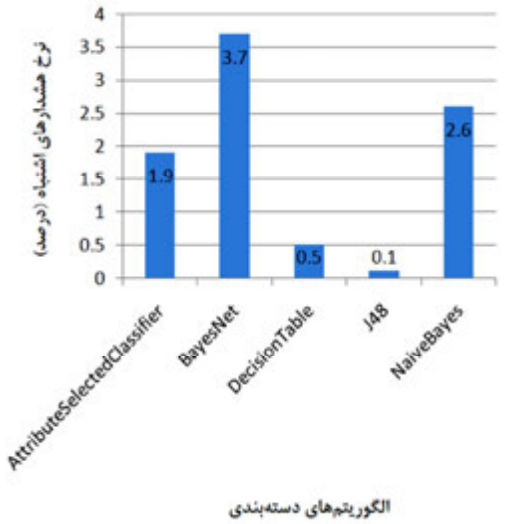


شکل 4: درصد تعداد نمونه‌هایی که در فاز آزمایش الگوریتم‌ها به درستی تشخیص داده شده‌اند.

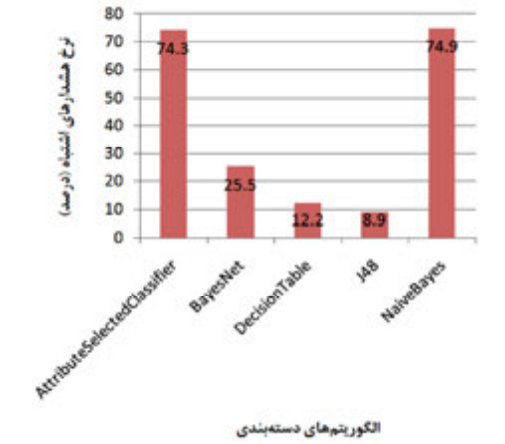
شکل 5 را مشاهده نمایید. نتایج آزمایشات همچنین نشان می‌دهد که در زمان آموزش دادن الگوریتم، الگوریتم C4.5 دارای نرخ هشدارهای اشتباه به مراتب کمتری نسبت به سایر الگوریتم‌ها می‌باشد و حتی می‌توان از این نرخ هشدار اشتباه نیز چشم‌پوشی کرد.

همچنین شکل (6) نشان می‌دهد که الگوریتم درخت تصمیم‌گیری C4.5 باز هم دارای نرخ هشدارهای اشتباه کمی می‌باشد. نرخ هشدارهای اشتباه الگوریتم‌هایی نظیر AttributeSelectedClassifier و NaiveBayes

می‌تواند ناشی از یادگیری بیش از حد[41] در فاز آموزششان باشد.



شکل (5): نرخ هشدارهای اشتباه در فاز آموزش الگوریتم‌ها



شکل (6): نرخ هشدارهای اشتباه در فاز آزمایش آلگوریتم‌ها

7. نتیجه‌گیری

همان‌گونه که در این مقاله عنوان شد، سیستم تشخیص نفوذ، با استفاده از یک‌سری قوانین از پیش تعریف‌شده دسترسی کاربران را بررسی و محدود می‌نماید. این قوانین، بر پایه‌ی دانش افراد مشخص شده است. این سیستم، تمام نفوذهای مهاجمان را مورد شناسایی قرار می‌دهد و فعالیت‌های لازم را برای متوقف نمودن حملات اتخاذ می‌کند. مسأله‌ی تشخیص نفوذ به طور وسیعی در زمینه‌ی امنیت شبکه‌های رایانه‌ای مورد مطالعه قرار است و اخیراً در زمینه‌های یادگیری ماشین و داده‌کاوی توجه زیادی به آن شده است.

در این مقاله به طور مختصر در مورد داده‌کاوی و یادگیری ماشین صحبت کردیم و به نحوه‌ی اِعمال این تکنیک‌ها به سیستم‌های تشخیص نفوذ اشاره کردیم. همچنین، از طریق شبیه‌سازی یکی از الگوریتم‌های معروف یادگیری ماشین، C4.5، نشان دادیم که این الگوریتم در قیاس با سایر الگوریتم‌های معروف، دارای نرخ کشف نفوذ به مراتب بالاتری است. برای کار آینده، بحث بر روی تکنیک‌های کاهش بُعد داده‌ها و انتخاب مؤثرترین ویژگی‌های مجموعه داده‌ها، می‌تواند در دستور کار پژوهشگران قرار بگیرد. این پژوهش می‌تواند افق روشنی را در باب کاهش زمان تشخیص نفوذ به سیستم‌ها و نیز افزایش جامعیت مجموعه داده‌ها بگشاید.

8. مراجع

[1] Lin, Ying and Zhang, Yan and Ou, Yang-jia. "*The design and implementation of host-based intrusion detection system*". In Proceedings of the 3rd International Symposium on Intelligent Information Technology and Security Informatics, pp. 595–598, 2010, April.

[2] Sabahi, F., &Movaghar, a.. "*Intrusion detection: A survey*'. In Proceedings of the 3rd International Conference on Systems and Networks Communications pp. 23–26, New York, NY, USA, 2008, October.

[3] Chang-tien Lu, Arnold P Boedihardjo, PrajwalManalwar and Falls Church. "*Exploiting Efficient Data Mining Techniques to Enhance Intrusion Detection Systems*".IRI 2005 IEEE International Conference on Information Reuse and Integration Conf 2005.

[4] http://www.kmining.com/info_definitions.html .

[5] Ming Xue, Changjun Zhu."*A Study and Application on Machine Learning of Artificial Intelligence*". International Joint Conference on Artificial Intelligence, pp. 272-274, 2009.

[6] S.Y. Wua and , E. Yen b, "*Data mining-based intrusion detectors*", Expert Systems with Applications: An International Journal Volume 36 Issue 3, pp. 5605–5612, 2009, April.

[7] http://www.cs.waikato.ac.nz/ml/weka/

[8] M. A. Hall, "Correlation-based Feature Subset Selection for Machine Learning. Hamilton", New Zealand, 1998.

[9] http://kdd.ics.uci.edu/databases/kddcup99/

زیرنویس

[1] Intrusion Detection Systems(IDS)
[2] Audit data
[3] Log
[4] Network based
[5] Promiscuous
[6] Header
[7] Host based
[8] Administrator
[9] Misuse detection
[10] Decision tree
[11] Genetic algorithm
[12] Fuzzy logic
[13] Clustering
[14] Bayesian network
[15] Data cleaning
[16] Data integration
[17] Data transformation
[18] Data mining
[19] Pattern evaluation
[20] Knowledge discovery
[21] Bayesian decision theory
[22] Multi-variate methods
[23] Classification trees
[24] Linear discrimination
[25] Multilayer perception
[26] Local models
[27] Hidden Markov models
[28] Reinforcement learning
[29] Flow chart
[30] Node splitting
[31] Entropy function
[32] Information gain
[33] Gini function
[34] Pre-pruning
[35] Post-pruning
[36] Information gain
[37] Basic features
[38] Content features
[39] Time-based traffic features
[40] Host-based traffic features
[41] Overfitting