dataacademy.ir

به کار گیری تکنیک های داده کاوی برای بهبود تشخیص نفوذ در شبکه های کامپیوتری

علی بازقندی [*,1]، مسعود خراسانیان [2]

۱- عضو هیات علمی دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی شاهرود، ایران

۲- دانشجوی کارشناسی، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی شاهرود، ایران

خلاصه

شبکه های کامپیوتری و اینترنت در معرض تعداد روز افزونی از حملات اینترنتی هستند. با انواع جدید حملات که به طور پیوسته آشکار می شوند، توسعهٔ رویکرد های سازگار و انعطاف پذیر براساس امنیت، یک چالش همیشگی است. در این زمینه تکنیک های تشخیص نفوذ در شبکه بر مبنای ناهنجاری، تکنولوژی ارزشمندی برای حفاظت از سیستم های مورد هدف (قرار گرفته) و شبکه ها در مقابل فعالیت های بدخواهانه است. تکنیک های داده کاوی متعددی به منظور تشخیص نفوذ در شبکه های کامپیوتری به کار گرفته شده اند. نتایج حاصل از ارزیابی ما نشان می دهد که در تشخیص نوع حملات PROBE, DOS , R2L تکنیک داده کاوی C4.5 عملکرد بهتری داشته است ولی در تشخیص حمله U2R روش AdaBoostM1 عملکرد قابل قبول تری دارد.

کلمات کلیدی: سیستم تشخیص نفوذ(IDS)، شبکه، حمله، ناهنجاری، داده کاوی،KDDCUP

۱. مقدمه

قبل از اینکه در مورد نفوذ در شبکه های کامپیوتری تو ضیح بدهیم. بهتر است بهتر است مطالبی را در مورد ناهنجاری شرح دهیم. نفوذ و ناهنجاری دو مقوله متفاوت می باشند. در حقیقت حالات مختلفی که این دو می توانند نسبت به هم داشته باشند را می توان به صورت شکل (۱) رسم کرد. اولین گام در تشخیص ناهنجاری، فرموله کردن مسأله (به شکل یک مسأله‌ی الگوشناسی) است. پس از آن، الگوریتم‌هایی که برای تشخیص ناهنجاری به کار گرفته می‌شوند و بهتر است به‌نحوی پیاده‌سازی شوند که توانایی یادگیری بر خط (online) را داشته باشند. مهم‌ترین مسأله در الگوریتم‌های تشخیص ناهنجاری توانایی به‌روز کردن نمایه‌ها یا الگوهای عادی به صورت پویا است. حالاتی را که شکل ۱ نمایش می دهد را در زیر خلاصه می کنیم:

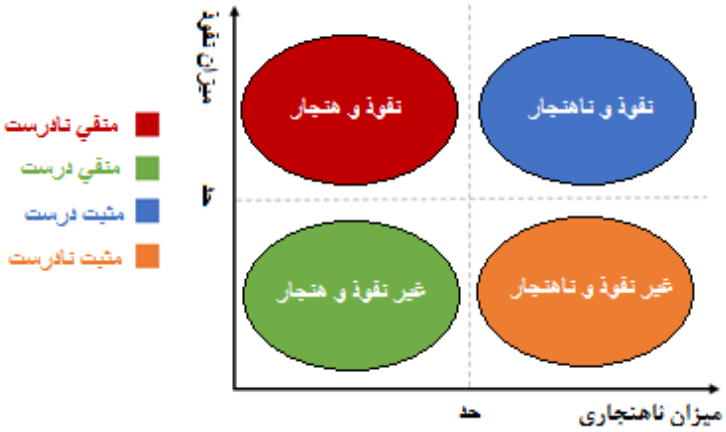
۱. نفوذ اما هنجار: یک سامانهٔ تشخیص ناهنجاری ممکن است در شناسایی این دسته از نفوذها ناموفق عمل کند. چرا که رفتاری شبیه به رفتاری عادی برای سامانه دارند (که توسط مهاجم اتخاذ شده است). یعنی یک سامانهٔ تشخیص ناهنجاری، به اشتباه خروجی منفی برای آن ثبت می‌کند.

* Corresponding author: Ali Bazghandi
Email: bazghandi@shahroodut.ac.ir

۲. غیرنفوذ اما ناهنجار:این موقعیت، همان موقعیتی است که یک سامانهٔ تشخیص ناهنجاری مثبت نادرست تولید می‌کند. اگرچه رفتار غیرعادی است، اما نفوذ نیست.

۳. غیرنفوذ و هنجار: موقعیتی که سامانه، منفی درست ایجاد می‌کند. هم رفتار عادی است و هم نفوذی صورت نگرفته است.

۴. نفوذ و ناهنجار: موقعیتی که سامانه، مثبت درست ایجاد می‌کند. یعنی رفتاری که شبیه به رفتار عادی سامانه نیست و یک حمله واقعاً صورت گرفته است. [۱]

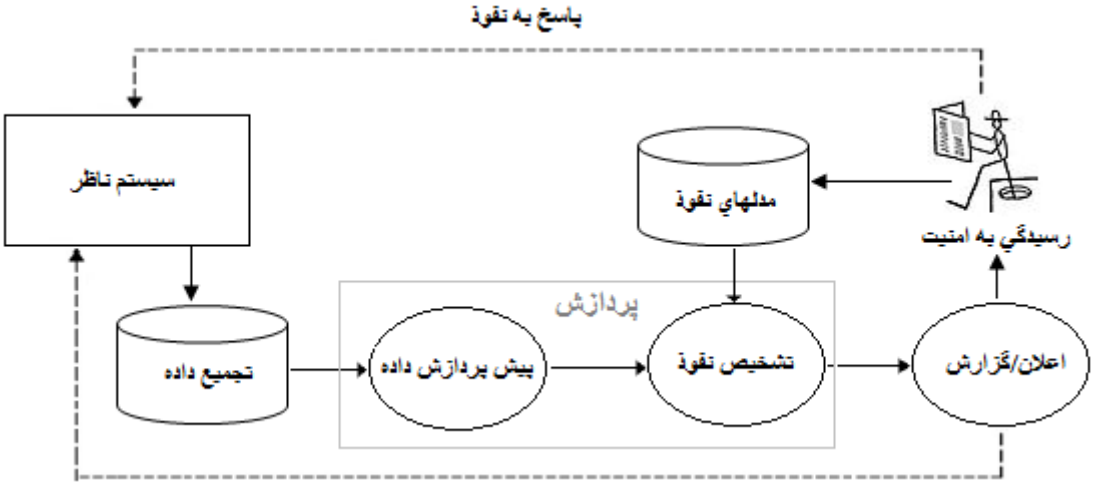


شکل ۱. رفتار ثبت شده از شبکه

همان‌طور که در دسته‌بندی موجود در تصویر مشخص است،در یک سامانه‌ی تشخیص ناهنجاری با دریافت هر رفتار شبکه، یک وضعیت برای آن ایجاد می‌شود، در صورتی که وضعیت مثبت درست یا منفی درست باشد، سامانه عملکرد صحیحی دارد و در غیر این‌صورت خطا وجود دارد. خطای مثبت نادرست مسئله‌ی بسیار مهمی در سامانه‌های تشخیص ناهنجاری محسوب می‌شود، این مسئله درست برخلاف سامانه‌های تشخیص نفوذی است که مسئله‌ی منفی نادرست در آن‌ها بیش‌تر دیده می‌شود.

آکادمی داده
dataacademy.ir

یک سیستم تشخیص نفوذ در حالت کلی به صورت شکل ۲ می باشد. [۲] سیستم ناظر همه بسته های وارده روی شبکه را دریافت می کند و آن را ذخیره می کند. در بخش پردازش با اطلاعات پیش پردازش شده و مدلهای نفوذ هشدار و اعلان فعال می شود. باز خوردهایی هم در ازای هشدار به صورت بی واسطه و با واسطه (انسان) به سیستم ناظر می گردد.



شکل ۲. سیستم تشخیص نفوذ در حالت کلی

در این مقاله در بخش ۲ در مورد داده کاوی، در بخش ۳ درباره مجموعه داده KDD CUP (Knowledge Discovery and Data Mining) توضیح داده می شود و در بخشهای ۴، ۵، ۶ و ۷ به ترتیب کارهای انجام شده، انتخاب ویژگی، روش پیشنهادی و نتایج آورده می شود.

۲. کامل داده کاوی

داده کاوی به بهره‌گیری از ابزار تجزیه و تحلیل داده، به منظور کشف الگوها و روابط معتبری که تا کنون ناشناخته بوده‌اند گفته می‌شود. این ابزارها ممکن است مدل های آماری، الگوریتم‌های ریاضی و روش‌های یاد گیرنده باشند که کار خود را به صورت خودکار و بر اساس تجربه‌ای که از طریق شبکه‌های عصبی یا درخت‌های تصمیم گیری به دست می‌آورند بهبود می‌بخشند. داده کاوی منحصر به گردآوری و مدیریت داده نبوده و تجزیه و تحلیل اطلاعات و پیش بینی را نیز شامل می‌شود. برنامه‌های کاربردی که با بررسی پرونده های متنی یا چندرسانه‌ای به کاوش داده می‌پردازند پارامترهای گوناگونی را در نظر می‌گیرد که عبارت اند از:

- قواعد انجمنی(Association) : الگوهایی که بر اساس آن یک رویداد به دیگری مربوط می‌شود مثلاً خرید قلم به خرید کاغذ.
- ترتیب(Sequence) : الگویی که به تجزیه و تحلیل توالی رویدادها پرداخته و مشخص می‌کند کدام رویداد، رویدادهای دیگری را در پی دارد مثلا تولد یک نوزاد و خرید پوشک.
- پیش بینی (Prediction) : در پیش بینی، هدف پیش بینی یک متغیر پیوسته می‌باشد. مانند پیش بینی نرخ ارز یا هزینه‌های درمانی.
- رده‌بندی یا طبقه‌بندی (Classification) : فرآیندی برای پیدا کردن مدلی است که رده‌های موجود در داده‌ها را تعریف می‌نماید و متمایز می‌سازد، با این هدف که بتوان از این مدل برای پیش بینی رده رکوردهایی که برچسب رده آنها (متغیر هدف)، ناشناخته می‌باشد، استفاده نمود. در حقیقت در رده‌بندی بر خلاف پیش بینی، هدف پیش بینی مقدار یک متغیر گسسته است. روش‌های مورد استفاده در پیش بینی و رده‌بندی عموما یکسان هستند.

dataacademy.ir

- خوشه بندی (Clustering) : گروه بندی مجموعه‌ای از اعضاء، رکوردها یا اشیاء به نحوی که اعضای موجود در یک خوشه بیشترین مشباهت را به یکدیگر و کمترین شباهت را به اعضای خوشه‌های دیگر داشته باشند.
- مصورسازی (visualization) : مصورسازی داده‌ها یکی از قدرتمندترین و جذاب ترین روش‌های اکتشاف در داده‌ها می‌باشد.

برنامه‌های کاربردی که در زمینه تجزیه و تحلیل اطلاعات به کار می‌روند از امکاناتی چون پرس و جوی ساخت یافته (Structured query)که در بسیاری از بانک‌های اطلاعاتی یافت می‌شود و از ابزارهای تجزیه و تحلیل آماری برخوردارند اما برنامه‌های مربوط به داده کاوی در عین برخورداری از این قابلیت‌ها از نظر نوع با آنها تفاوت دارند. بسیاری از ابزار ساده برای تجزیه و تحلیل داده، روشی بر پایه راستی آزمایی(verification) را به کار می‌برند که در آن فرضیه‌ای بسط داده شده و آنگاه داده‌ها برای تایید یا رد آن بررسی می‌شوند. به طور مثال ممکن است این نظریه مطرح شود که فردی که یک چکش خریده حتماً یک بسته میخ هم خواهد خرید. کارایی این روش به میزان خلاقیت کاربر برای ارایه فرضیه‌های متنوع و همچنین ساختار برنامه به کار رفته بستگی دارد. در مقابل در داده کاوی روش هایی برای کشف روابط به کار برده می‌شوند و به کمک الگوریتم‌هایی روابط چند بعدی بین داده‌ها تشخیص داده شده و آنهایی که یکتا (unique)یا رایج هستند شناسایی می‌شوند. به طور مثال در یک فروشگاه سخت‌افزار ممکن است بین خرید ابزار توسط مشتریان با تملک خانه شخصی یا نوع خودرو، سن، شغل، میزان درآمد یا فاصله محل اقامت آنها با فروشگاه رابطه‌ای برقرار شود.

در نتیجه با توجه به قابلیت‌های پیچیده‌ای که دارد، برای موفقیت در تمرین داده کاوی دو مقدمه مهم است یکی فرمول واضحی از مشکل که قابل حل باشد و دیگری دسترسی به داده متناسب. بعضی از ناظران داده کاوی را مرحله‌ای در روند کشف دانش در پایگاه داده‌ها می‌دانند. مراحل دیگری در روندKDD ، به صورت تساعدی شامل، پاکسازی داده، انتخاب داده انتقال داده، داده کاوی، الگوی ارزیابی، و عرضه دانش می‌باشد. بسیاری از پیشرفت‌ها در تکنولوژی و فرایندهای تجاری بر رشد علاقه‌مندی به داده کاوی در بخش‌های خصوصی و عمومی سهمی داشته‌اند. بعضی از این تغییرات شامل:

- رشد شبکه‌های کامپیوتری که برای ارتباط برقرار کردن پایگاه های داده مورد استفاده قرار می‌گیرند.
- توسعه و افزایش تکنیک هایی بر پایه جستجو مثل شبکه‌های عصبی و الگوریتم‌های پیشرفته.
- گسترش مدل محاسبه کارگزار-مشتری که به کاربران اجازه دسترسی به منابع داده‌های متمرکز شده را از روی میز کاری می‌دهد.
- و افزایش توانایی به ادغام داده از منابع غیر متناجس، به یک منبع قابل جستجو می‌باشد.

علاوه بر پیشرفت ابزارهای مدیریت داده، افزایش قابلیت دسترسی به داده و کاهش نرخ نگهداری داده نیز نقش ایفا می‌کند. در طول چند سال گذشته افزایش سریع جمع‌آوری و نگهداری حجم اطلاعات وجود داشته‌است. با پیشنهادهای برخی از ناظران مبنی بر آنکه کمیت داده‌های دنیا به طور تخمینی هر ساله دوبرابر می‌گردد. در همین زمان هزینه ذخیره‌سازی داده‌ها به طور قابل توجهی کاهش پیدا کرده‌است. پیرو آن قدرت محاسبه‌ها در هر ۱۸ تا ۲۴ ماه به دوبرابر ارتقاء پیدا کرده‌است این در حالی است که هزینه قدرت محاسبه رو به کاهش است. داده کاوی به طور معمول در دو حوزه خصوصی و عمومی افزایش پیدا کرده‌است. سازمان ها، داده کاوی را به عنوان ابزاری برای بازدید اطلاعات مشتریان، کاهش تقلب و اتلاف و کمک به تحقیقات پزشکی استفاده می‌کنند. با این همه ازدیاد داده کاوی به تبع، بعضی از پیاده‌سازی ها و پیامدهای اشتباه را هم در پی دارد. این ها شامل نگرانی‌هایی در مورد کیفیت داده‌ای که تحلیل می‌گردد، توانایی کار

گروهی پایگاه های داده و نرم‌افزارهای بین سازمان ها و تخطی‌های بالقوه به حریم شخصی می‌باشد. همچنین ملاحظات در مورد محدودیت داده کاوی در سازمانهایی که کارشان تاثیر بر امنیت است، نادیده گرفته می‌شود. [۳]

۳. مجموعه داده KDDCUP99

گروه Integrated Service Technology(IST) از آزمایشگاه MIT Lincoln زیر نظر Defense Advanced Research Projects Agency (DARPA) و Air Force Research Laboratory (AFRL) اولین داده های استاندارد را برای بررسی و ارزیابی سیستم های تشخیص نفوذ جمع آوری نمودند. این اطلاعات در طول چند هفته در یک شبیه سازی برای آزمایش سیستم تشخیص نفوذ DARPA به کار رفته اند. این مجموعه داده ها براساس سال جمع آوری اطلاعات (1998-1999-2000) دسته بندی شده اند.

مجموعه داده های سال 1999 که به اهتمام و با نظارت لی و در طی انجام پروژه دکترای وی جمع آوری شد، در سومین مسابقه بین المللی اکتشاف دانش و داده کاوی، KDDCPU 99 و در پنجمین کنفرانس در این زمینه مورد استفاده قرار گرفت. این بانک اطلاعاتی شامل رکوردهای اتصال استانداردی می باشد که مجموعه ای از حملات و نفوذهای شبیه سازی شده در یک شبکه نظامی را شامل می شود.

یک اتصال، دنباله ای از بسته های با پروتکل User Datagram Protocol(UDP)، Transmission Control Protocol(TCP) و یا Internet Control Message Protocol(ICMP) است که در زمان های مشخص شروع و پایان می یابد و بین آن متن ها، داده ها از آدرس IP مبدا به آدرس IP مقصد و برعکس، تحت یک قرارداد تعریف شده جریان دارند. هر اتصال به عنوان نرمال یا حمله برچسب گذاری می شود و در مورد حمله، نوع آن دقیقا مشخص می گردد. رکورد هر اتصال شامل حدود صد بایت است. حملاتی که در این مجموعه داده مشاهده می شوند، در چهار دسته اصلی دسته بندی می شود:

- DOS (Denial Of Service): در این نوع حمله، مهاجم تعداد زیادی درخواست به یک میزبان ارسال می کند.
- R2L (Remote to Local): در این نوع حمله، مهاجم سعی دارد که با استفاده از نقاط آسیب پذیر سیستم، کنترل یک ماشین خارجی را از طریق شبکه به عنوان یک کاربر محلی بدست آورد.
- U2R (User to Root): در این نوع حمله، مهاجم سعی دارد از یک ماشین خارجی دسترسی غیرمجازی فرضا برای دستیابی به Root سیستم پیدا کند.
- Probing: در این نوع حمله مهاجم سعی دارد به منظور تجسس، اطلاعاتی را در مورد ماشین ها و سرویس های شبکه به دست آورد.

۴. کارهای انجام شده

در سال ۱۹۸۰ مفهوم سیستم تشخیص نفوذ (IDS) با مقاله اولیه اندرسون آغاز شد. با معرفی این اندیشه که پیوست ها (که اغلب جهت ممیزی استفاده می شوند)، دارای اطلاعاتی حیاتی هستند که ممکن است در تعقیب رفتارهای غیرعادی و درک رفتارهای کاربر مفید باشد. در واقع کار او آغاز IDS های برمبنای میزبان بود. در سال ۱۹۸۶ دکتر دوروتی دنینگ مدلی منتشر کرد که اطلاعات ضروری برای توسعهٔ IDS های تجاری را آشکار نمود. [۴] MIDAS (Modular Intelligence Data Analysis System) یک سیستم خبره با استفاده از P-Best و LISP در سال

۱۹۸۸ پیاده سازی شد. در همان سال Haystack نیز پیاده سازی شد که با استفاده از آمار سعی در کم کردن پیوست های ممیزی داشت. Wisdom & Sence در سال ۱۹۸۹ به عنوان یک تشخیص دهنده ی ناهنجاری مبتنی برآمار که براساس تحلیل آماری، قوانینی تولید می کرد و سپس از آن قوانین برای تشخیص ناهنجاری استفاده می کرد، پیاده سازی شد. Heberlein در سال ۱۹۹۰ ابتدا ایده IDS های شبکه، توسعه نظارت (Monitoring) امنیت شبکه و IDS های ترکیبی را مطرح کرد و پس از او Lunt یک سیستم خبره تشخیص نفوذ به نام SRI را ارائه کرد، سیستمی با دو رویکرد: یک سیستم خبره مبتنی بر قانون و یک تشخیص دهنده ناهنجاری بر مبنای آمار که بر روی پایگاه های کاری شرکت سان (Sun) اجرا شد و قادر بود داده را هم در سطح کاربر و هم در سطح شبکه مورد بررسی قرار دهد.[۵] از سوی دیگر در اوایل دهۀ ۹۰ توسعۀ تجاری IDS ها آغاز شد و ماشین استنتاجی مبتنی بر زمان، TIM (Time based Inductive Machine)، با استفاده از یادگیری استنتاجی الگوهای متوالی و مشترک کاربر در LISP، بر روی یک کامپیوتر VAX 3500 ناهنجاری را تشخیص می داد. در سال ۱۹۹۱ IDS های توزیع شده (Distributed IDS) شامل یک سیستم خبره که توسط محققین دانشگاه کالیفرنیا ساخته شد، یک تشخیص دهنده ناهنجاری مبتنی بر آمار به نام NADIR و یک سیستم خبره توسط Los Alamos National Laboratory's Integrated Computing Network پیاده سازی شدند. Lunt در سال ۱۹۹۳ نسل دوم سیستم خبره تشخیص نفوذ را با توسعه SRI با استفاده از شبکه عصبی مصنوعی ارائه کرد. در سال ۱۹۹۸ آزمایشگاه ملی Lawrence Berkely یک زبان قانون نویسی به نام Bro را برای تحلیل بسته ها از مجموعه داده libpcap معرفی کرد. در سال ۲۰۰۱ در تحلیل داده های ممیزی، IDS های کاوشگر از tcpdump برای ایجاد پروفایل های قوانین (برای طبقه بندی) استفاده شد.[۵] اکنون به توضیح مختصری از کارهای جدید چندسال اخیر می پردازیم:

- در کاری که Dewam Md. Farid و Mohammad Zaridur Rahman در سال ۲۰۱۰ انجام داده اند با ارائه الگوریتم جدیدی با عنوان الگوریتم Bayesian (وفقی خود بهبود دهنده)، رویکرد جدیدی در طبقه بندی هشدارها اتخاذ کرده اند که تعداد مثبت های نادرست یعنی رفتارهای عادی که به اشتباه به عنوان رفتار ناهنجار اعلام شده و هشدار داده می شوند، را پایین می آورد و قدرت تشخیص درست را بالا می برد. رویکرد ارائه شده بر روی دامنه امنیت تشخیص نفوذ در شبکه مبتنی بر ناهنجاری اجرا شده و توانسته است به درستی انواع مختلف حمله را در مجموعه داده KDD99 با نرخ بالایی در زمان کم، طبقه بندی کند و با استفاده از منابع محاسباتی محدود تعداد مثبت های نادرست را کاهش دهد. [۶]
- در کاری که Mohammad Sazzadul Hoque ، Md. Abdul Mukit و Md. Abu Naser Bikas در سال ۲۰۱۲ انجام داده اند سیستم تشخیص نفوذی با استفاده از الگوریتم ژنتیک برای تشخیص کارآمد انواع نفوذ در شبکه ارائه کردند. در این رویکرد از تئوری تکامل برای تکامل اطلاعات جهت فیلتر کردن داده های ترافیک شبکه و کم کردن پیچیدگی استفاده شده است. برای پیاده سازی و اندازه گیری کارایی سیستم از مجموعه داده KDD99 استفاده شده که در این شرایط نرخ خوبی در تشخیص برای خود ثبت کرده است. [۷]
- در کاری که R. Remya و Amrita Vishwa Vidyapeetham در سال ۲۰۱۳ انجام داده اند از الگوریتم ژنتیک و SOFM (Self-Organizing Feature Map) برای بهبود تشخیص ناهنجاری و کم کردن خطای تشخیص استفاده شد. در این کار تشخیص ناهنجاری توسط یک ماشین برداری پشتیبانی (SVM) با حاشیۀ نرم صورت می گیرد که ورودی ها را با توجه به رفتارشان به دو دسته عادی و ناهنجار تقسیم می کند. الگوریتم های ژنتیک و SOFM برای بهبود ویژگی ها و استخراج اطلاعات از یک مجموعه داده بزرگ مثل KDD99 استفاده می شوند. الگوریتم ژنتیک کمک بزرگی در تشخیص رفتار ناهنجار می کند و SOFM کمک می کند تا گروه های مشابه از داخل مجموعه داده به وسیلۀ اندازه گیری میزان تشابه، احراز هویت گردند. این دو الگوریتم

یادگیری ماشین، باعث کاهش حجم مجموعهٔ داده و ویژگی ها در یادگیری SVM می شوند. چارچوب ارائه شده با نام GSS (Generic Security Services) دارای ۱۰ درصد افزایش در نرخ تشخیص و ۵۰ درصد کاهش در نرخ خطاهای تشخیص نسبت به SVM با حاشیه نرم است.

۵. روش پیشنهادی

ما در این بررسی، میزان دقت در صحت تشخیص حمله در مجموعه داده KDD cup 99 را از طریق سه الگوریتم RandomTree , AdaBoostM1+RandomTree و J48 را به کمک نرم افزار Weka ارزیابی و مقایسه نماییم. به همین منظور ابتدا یکبار همه ۴۲ مشخصه مجموعه داده را در نرم افزار Weka انتخاب و سه الگوریتم ذکر شده را روی آن اعمال و نتایج حاصل را با یکدیگر مقایسه می نماییم. (جدول ۱)

جدول ۱- میزان دقت در صحت تشخیص حمله (با تمام ۴۲ مشخصه)

نوع حمله	نرخ نمونه در مجموعهٔ آزمون	J48	RandomTree	AdaBoostM1+ RT
U2R	52	54.80%	45.11%	53.15%
R2L	1126	97.23%	92.67%	94.28%
PROBE	4107	98.65%	98.58%	98.57%
Dos	391458	99.99%	99.99%	99.99%
normal	97278	99.90%	99.90%	99.90%

همان‌طور که مشاهده می‌شود، درخت تصمیم J48 در تمام حمله ها نسبت به دو روش دیگر موفق تر عمل کرده است و در تشخیص حمله از دقت بالاتری برخوردار بوده است. اما نتایج نشان می دهد که در میان انواع حمله در نوع U2R دقت تشخیص تقریبا پایین می باشد. حال ما قصد داریم بررسی کنیم که با انتخاب ۱۲ مشخصه ای که بهره اطلاعاتی (Gain) بالاتری نسبت به سایر مشخصه ها دارند می توانیم دقت تشخیص را در این حمله افزایش دهیم یا خیر؟ و اینکه آیا با این کار باز هم الگوریتم J48 از دو روش دیگر بهتر خواهد بود؟

۱.۵ بهره اطلاعاتی (Information Gain)

محاسبه بهره اطلاعاتی نسبتاً ساده است، در این روش ابتدا با داشتن تمام مجموعه داده‌ی آزمایشی، با داشتن برچسب کلاس، تعداد برچسب هر کلاس را می‌شماریم و مطابق فرمول زیر محتوای اطلاعات کل این مجموعه‌داده را محاسبه می‌کنیم:

$$Info(D) = -\left[\frac{\#C1\ \ Labled}{\#all\ \ Instances}\right] * \log\frac{\#C1\ \ Labled}{\#all\ \ Instances} - \left[\frac{\#C2\ \ Labled}{\#all\ \ Instances}\right] * \log\frac{\#C2\ \ Labled}{\#all\ \ Instances} \qquad (1)$$

گفتنی است اگرچه در مجموعه‌داده‌ی KDD99 چندین نوع حمله به تفکیک مشخص شده است، اما در این نوع روش تشخیص دو کلاس هنجار و ناهنجار وجود دارد، بنابراین به دست آوردن اطلاعات بسیار ساده خواهد شد. در مرحله‌ی بعدی به ازای هر ویژگی و بر اساس مقادیر مجاز تعریف‌شده، (مانند فرمول مطرح‌شده در بالا، با این تفاوت که به جای هر کلاس، مقادیر مجاز برای ویژگی درج می‌شود)، مقدار اطلاعاتی که هر ویژگی تولید می‌کند را محاسبه می‌کنیم، سپس مقدار بهره ی اطلاعاتی هر ویژگی از فرمول زیر محاسبه می گردد:

$$Gain(f) = Info(D) - \inf o_f1(D) \qquad (2)$$

۲.۵ انتخاب ویژگی بر اساس بهره اطلاعاتی

همان طور که در بالا توضیح داده شد مقدار بهره اطلاعاتی (gain) برای ویژگی‌ها مشخص می‌شود و ویژگی های برتر انتخاب می‌شوند. (جدول ۲) گفتنی است ویژگی سرویس که ساخت درخت بر اساس آن شکل می‌گیرد، بهره بالایی دارد، و نشان می‌دهد رده‌بندی داده‌ها بر اساس این ویژگی بی‌اساس نیست.

جدول ۲- مقدار بهره ۱۲ مشخصه برتر مجموعه داده

ردیف	ویژگی	بهره
۱	src_byte	0.939935
۲	Service	0.832597
۳	Count	0.807751
۴	dst_byte	0.781945
۵	logged_in	0.582982
۶	dst_host_srv_diff_host_rate	0.441058
۷	dst_host_diff_srv_rate	0.421889
۸	dst_host_count	0.404109
۹	srv_count	0.365569
۱۰	Flag	0.328112
۱۱	dst_host_serror_rate	0.306328
۱۲	dst_host_srv_serror_rate	0.304958

چنان که گفته شد از میان ۴۲ مشخصه مجموعه داده، ۱۲ مشخصه ای که بهره بالاتری داشتند را جدا می نماییم و مجدد سه الگوریتم یاد شده (AdaBoostM1+RandomTree , RandomTree و J48) را آزمایش می نماییم. نتایج حاصل از آزمایش را در جدول زیر (جدول ۳) مشاهده می نمایید.

جدول ۳- میزان دقت در صحت تشخیص حمله (با ۱۲ مشخصه)

نوع حمله	نرخ نمونه در مجموعهٔ آزمون	J48	RandomTree	AdaBoostM1+ RT
U2R	52	45.23%	60.35%	69.23%
R2L	1126	97.21%	95.49%	94.04%
PROBE	4107	98.46%	98.24%	98.48%
Dos	391458	99.99%	99.96%	99.96%
normal	97278	99.90%	99.90%	99.90%

۶. نتیجه گیری

با بررسی نتایج بدست آمده از جدول ۲ و ۳ در بخش قبل درمی یابیم که اعمال روش بهره اطلاعاتی در تعیین مشخصه و سپس استفاده از الگوریتم AdaBoostM1+ RT تاثیر خوبی داشته و باعث شده است که دقت تشخیص حمله در حمله نوع U2R تا مقدار قابل قبولی افزایش یابد. پس با توجه به آزمایش های صورت گرفته و نتایج حاصل از آن پیشنهاد می گردد که در تشخیص حملات فقط از یک الگوریتم به تنهایی استفاده نگردد، بلکه برای هر نوع حمله از الگوریتمی استفاده شود که نتایج مطلوب تری به ما می دهد. در این آزمایش و با ارزیابی و مقایسه سه الگوریتمJ48,RandomTree, AdaBoostM1+RandomTree دریافتیم که اگر در تشخیص نوع حمله U2R از الگوریتم AdaBoostM1+ RT و با اعمال مولفه ی بهره استفاده نماییم به نتایج بهتر و مطلوب تری دست خواهیم یافت. اما برای سایر حمله ها الگوریتم J48 نتایج بهتری نسبت به دو الگوریتم دیگر به ما می دهد.

۷. مراجع

1.Monowar Hussain Bhuyan1, D K Bhattacharyya1 and J K Kalita2. "*Survey on Incremental Approaches for Network Anomaly Detection*", International Journal of Communication Networks and Information Security (IJCNIS),Vol. 3, No. 3,(2011).

2.G.V. Nadiammai, M. Hemalatha, "*Effective approach toward Intrusion Detection System using data mining techniques*", Egyptian Informatics Journal, 2013

3.K. Bharat, B. Chang, M. Henzinger, M. Ruhl. "*Who links to whom: Mining linkage between web sites*", IEEE International Conference on Data Mining (ICDM '01), SanJose, California, November 2001.

4.Denning ED. "*An intrusion-detection model*", IEEE Transactions on Software Engineering 1987;13(2):222–32.

5.Dewan Md. Farid , Mohammad Zahidur Rahman. "*Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm*", Journal of Computers, Vol. 5, No. 1, January 2010

6.Dewan Md. Farid and M. ZahidurRahman, "*Anomaly network intrusion detection based on improved self adaptive Bayesian algorithm*", Journal of Computers, Academy publisher, Vol. 5, No. 1, January 2010, pp 23-31.

7.Mohammad Sazzadul Hoque, Md Abu Naser Bikas, Abdul Mukit, "*An Implementation of Intrusion Detection System using Genetic Algorithm"*, International Journal of Network Security & Its Applications (IJNSA), March 2012.

آکادمی داده
dataacademy.ir