استفاده از الگوریتم کلونی مورچگان و روش یادگیری زوجی جهت طبقه‌بندی حملات در سیستم‌های تشخیص نفوذ

محمد علی ندومی[۱]، مجید سینا[۲]

۱. کارشناس ارشد، تحصیلات تکمیلی، دانشگاه آزاد اسلامی، بوشهر، ایران nadoomi@gmail.com
۲. دکترا، تحصیلات تکمیلی، دانشگاه آزاد اسلامی، بوشهر، ایران majidsina.edu@gmail.com

چکیده

سیستم‌های تشخیص نفوذ برای ایجاد امنیت در شبکه‌های کامپیوتری پیشنهاد شده‌اند تا در صورتی که نفوذگر از سایر تجهیزات امنیتی عبور کرد، بتواند آن را تشخیص داده و از پیش‌روی آن جلوگیری کند. یکی از مهمترین چالش‌های این سیستم‌ها، ابعاد بالای داده‌های آن می‌باشد. در این تحقیق برای کاهش ابعاد داده‌ای از یک الگوریتم ژنتیک ساده با طول رشته متغیر استفاده می‌کنیم. در مرحله بعد با توجه به ویژگی‌های انتخاب شده، یک مدل فراابتکاری جهت طبقه‌بندی داده‌ها، با استفاده از الگوریتم مورچه‌ها ارائه می‌دهیم. مدل طبقه‌بندی پیشنهادی سعی در تقسیم‌بندی داده‌ها به دو بخش نمونه‌های هنجیده و ناهنجیده دارد. جهت ارزیابی عملکرد روش پیشنهادی از پایگاه داده NSL-KDD که نسبت به سایر داده‌های تشخیص نفوذ از رکوردهای واقعی‌تری برخورد دار است، استفاده می‌کنیم. نتایج حاصل از آزمایشات، عملکرد بهتر روش پیشنهادی را در مقایسه با سایر روش‌های موجود نشان می‌دهد.

کلید واژه‌ها: انتخاب ویژگی، طبقه‌بندی، الگوریتم ژنتیک، الگوریتم مورچگان، پایگاه داده NSL-KDD

۱- مقدمه

نفوذ به مجموعه‌ای از اقدامات غیرقانونی که امنیت و محرمانگی یک منبع را به خطر بیاندازد، گفته می‌شود(۱). تشخیص و جلوگیری از نفوذ (IDS)[1] امروزه به عنوان یکی از مکانیزم های اصلی در برآوردن امنیت شبکه ها و سیستم های کامپیوتری مطرح است. زمانی که برای اولین بار ایده استفاده از این سیستم‌ها مطرح گردید، به دلیل بار پردازشی فراوان، تنها مورد استقبال محیط های نظامی و تجاری مهم قرار گرفت. با پیشرفت چشمگیر در طراحی و تولید سخت افزارها و توسعه معماری های نوین در طراحی و تولید نرم-افزارها، امکان استفاده از این ایده و تکنولوژی برای طیف گسترده ای از سیستم های کامپیوتری امکان پذیر شده است(۲). امروزه تشخیص، از فرآیندهای اصلی در برآوردن امنیت سیستم‌ها می‌باشد. زمانی که ایده اولیه سیستم‌های IDS مطرح گردید، تعریف جامعی از نفوذ مدنظر بود که هرگونه فعالیت در راستای مخدوش نمودن پایه‌های اصلی امنیت اطلاعات، محرمانگی، جامعیت و دسترس‌پذیری را در بر می‌گرفت. به دلیل محدودیت‌هایی که در پیاده‌سازی سیستم‌های اولیه وجود داشت، سیستم‌های تشخیص نفوذ، محدود به ناظرین فعالیت‌های سیستم‌عامل و ترافیک شبکه شدند و سیستم‌هایی با قابلیت‌های در حد تعریف اصلی، به دلیل هزینه‌های بسیار زیاد در طراحی و پیاده‌سازی، تنها در بسترهای خاص نظامی و اطلاعاتی مورد استفاده قرار گرفتند(۳).

تشخیص نفوذ، بخشی ضروری از یک سیاست امنیتی کامل در سیستم های اطلاعاتی است. از آنجایی که تعداد عظیمی از نفوذهای بالقوه در هر روز رخ می دهد، سیستمهای تشخیص نفوذ به منظور شناسایی این حملات سایبری توسعه یافته است. در ارتباط با

[1] Intrusion Detection System

مکانیزم های بازرسی امنیت که به صورت پویا، ورودهای صورت گرفته به سیستم و ترافیک شبکه را به منظور کسب اطلاع از سیستم مورد کنترل قرار می‌دهند، سیستم تشخیص نفوذ دارای تابعی برای تحلیل این اطلاعات بوده و سپس از الگوریتم های تشخیص نفوذ، به منظور تعیین اینکه آیا این رخدادها نشان دهنده حمله بوده یا نشان دهنده استفاده قانونی از سیستم، استفاده می کند. از شبکه‌های کامپیوتری برای اشتراک‌گذاری و ارتباط کاربران با یکدیگر و دسترسی به منابع اطلاعاتی استفاده می‌شود. هر فرد با توجه به سطح مجاز می‌تواند از شبکه و اطلاعات آن بهره ببرد(٤). اما همیشه تعدادی سودجو وجود داشته که به دنبال دسترسی غیرمجاز به شبکه می باشند تا از منابع آن جهت پیشبرد اهداف خصمانه بهره ببرند. سیستم‌های تشخیص نفوذ به منظور جلوگیری از این واقعه در شبکه های کامپیوتری لازم و ضروری می باشد.

وجود هشدارهای کاذب و همچنین خطای تشخیص، دو عامل اصلی است که سیستم های تشخیص به دنبال بهبود آن می‌باشند. در این پژوهش با هدف قرار دادن این دو عامل سعی می شود تا با سیستم یادگیری زوجی و بهبود آن توسط الگوریتم مورچگان سیستمی طراحی کنیم که علاوه بر کاهش نرخ هشدارهای کاذب دارای خطای تشخیص پایینی باشد. در عمل سیستم های امنیتی شبکه ممکن است از نظر ساختار و طراحی سخت افزاری به نحوی نباشند که بتوانند از کلیه نفوذها جلوگیری به عمل آورند. علاوه بر این، در مورد شبکه های پیاده سازی شده به دلیل پیشرفت انواع حملات و عدم به روز رسانی سیستم های سخت افزاری امنیت شبکه لازم است تا در عمل، یک سیستم تشخیص نرم افزاری یا همان سیستم تشخیص نفوذ که قادر به همگام شدن با انواع حملات جدید می باشد طراحی گردد تا بتواند از انواع حملات جلوگیری به عمل آورد(٥).

در ادامه این تحقیق به بررسی برخی از جدیدترین کارهای انجام شده در بخش ٢ می‌پردازیم، در بخش ٣ سیستم تشخیص نفوذ پیشنهادی مبتنی بر الگوریتم کلونی مورچه و روش یادگیری زوجی مطرح شده و عملگرهای لازم جهت افزایش دقت تشخیص ارائه می‌شود. نتایج حاصل از ارزیابی سیستم تشخیص نفوذ پیشنهادی در بخش ٤ آورده شده و در نهایت نتیجه‌گیری و پیشنهادات در بخش ٥ ذکر شده است.

٢- مروری بر تحقیقات انجام شده

تحقیقات بسیاری در زمینه روش‌های کاهش ویژگی و طبقه‌بندی سیستم‌های تشخیص نفوذ انجام شده است که در ادامه به معرفی چند تحقیق مشابه در این خصوص می‌پردازیم.

گویال و همکاران(٦)، یک روش مبتنی بر الگوریتم ژنتیک برای تشخیص نفوذ ارائه دادند که از یک رویکرد یادگیری ماشین برای انتخاب ویژگی‌ها استفاده می‌کند(٦). با تولید یک سری قوانین و بر اساس هر قاعده یک نوع خاص نفوذ شناسایی می‌شود. در تحقیق دیگر مودا و همکاران(٧)، رویکرد یادگیری ترکیبی خوشه‌بندی k-means و طبقه‌بندی بیز ساده را برای تشخیص نفوذ پیشنهاد دادند(٧). خوشه‌بندی تمام داده‌ها را به گروه‌های مربوطه قبل از استفاده طبقه‌بندی نسبت می‌دهد که این روش عملکرد مناسبی از جهت دقت و سرعت را حاصل نموده است. ساها و همکاران(٨)، یک سیستم تشخیص نفوذ مبتنی بر یادگیری ماشین توسعه دادند(٨). آنها الگوریتم ژنتیک را به همراه SVM برای تعیین اتوماتیک مجموعه مناسب از ویژگی‌ها به کار بردند. در این تحقیق یک دیکشنری از پیش تعریف شده از انواع حملات وجود دارد که این روش متمایزترین ویژگی‌ها برای هر نوع حمله را در دسترس قرار می‌دهد.

چا و همکاران(٩)، یک روش جدید برای انتخاب ویژگی‌های موثر در ساخت مدل تشخیص نفوذ ارائه دادند(٩). روش انتخاب ویژگی پیشنهادی با توجه به میانگین ویژگی‌ها از هر کلاس نسبت به کلاس کل محاسبه می‌شود. برای ارزیابی عملکرد روش پیشنهادی از روش‌های استاندارد مبتنی بر همبستگی[2]، مبتنی بر اطلاعات[3] و مبتنی بر وزن[4] استفاده شده است. بنایچا و همکاران(١٠)،

[2] Correlation

نیز سیستم تشخیص نفوذی با استفاده از الگوریتم ژنتیک توسعه دادند که در آن ایده‌ای برای بهبود بخش‌های ایجاد جمعیت اولیه و اپراتور انتخاب در الگوریتم ژنتیک ارائه شده است(۱۰).

در تحقیقی دیگر اقدم و کبیری(۱۱)، از الگوریتم کلونی مورچه به منظور استخراج ویژگی در سیستم تشخیص نفوذ استفاده کرده‌اند(۱۱). این روش به دلیل استفاده از زیر مجموعه‌های ساده برای کلاس‌بندی، قابلیت اجرای سریع و دارای پیچیدگی محاسباتی کمی می‌باشد. کاهش تعداد ویژگی‌ها با استفاده از نمایش گراف و اطلاعات اکتشافی برای به روزرسانی فرمون‌ها باعث ارائه دقت بالاتر در تشخیص نفوذ و هشدار اشتباه پایین‌تر شده است. موباراک(۱۲)، به بررسی روش‌های SVM و شبکه‌های عصبی SOM در سیستم تشخیص نفوذ پرداخته است(۱۲). تجزیه و تحلیل روش‌های مذکور روی دو مجموعه داده‌های تشخیص نفوذ نشان می‌دهد که روش SVM از کارایی و سرعت محاسباتی بالاتری نسبت به SOM برخودار می‌باشد. یکی از دلایل ضعف شبکه‌های عصبی در سیستم های تشخیص نفوذ مشکل تشخیص اندازه مناسب شبکه و وزن‌های آن می‌باشد. نتایج آزمایشات دقت بالای ۹۹٪ را برای روش SVM نشان می‌دهد.

در تحقیق دیگری کوریس و همکاران(۱۳)، یک مدل طبقه‌بندی ترکیبی را براساس الگوریتم‌های درختی برای تشخیص اختلال در شبکه مطرح کردند(۱۳). هدف این الگوریتم مشخص کردن این مسئله بود که آیا ترافیک‌های شبکه ورودی مبتنی بر ۴۱ ویژگی از الگوی ترافیک شبکه طبیعی الهام گرفته‌اند یا یک حمله محسوب می‌شوند. در این تحقیق از یک الگوریتم ترکیبی متشکل از درخت تصمیم و نایو بیز استفاده شده است. چن و همکاران(۱۴)، جهت‌گیری به سمت ایجاد یک طبقه‌بندی با استفاده از سیستم ایمنی مصنوعی (AIS) همراه با یادگیری افزایشی جمعیت محور (PBIL) و فیلترینگ مشارکتی را برای تشخیص نفوذ در شبکه مطرح ساختند(۱۴). AIS یک ابزار قدرتمند از نظر نابودی آنتی ژن‌ها می‌باشد و از پروسه سیستم‌های ایمنی طبیعی الهام گرفته است. PBIL از تجارب گذشته در ایجاد و تکامل گونه‌های جدید بواسطه یادگیری و تطبیق ایده فیلترینگ مشارکتی برای طبقه‌بندی بهره می‌برد.

وارما و همکاران(۱۵)، یک الگوریتم انتخاب ویژگی با استفاده از آنتروپی فازی نسبی و بهینه‌سازی مورچگان را برای یک سیستم تشخیص نفوذ واقعی ارائه دادند(۱۵). ایده اصلی این تحقیق انتخاب ویژگی‌های مبتنی بر زمان واقعی به منظور کاهش ترافیک شبکه می‌باشد. نتایج حاصل از آزمایشات دقت ۹۶٪ را در حالت دو کلاسه به ثبت رسانیده است. راوات و چوبیی(۱۶)، با استفاده از الگوریتم بهینه‌سازی مورچه‌ها یک سیستم تشخیص نفوذ جدید بر اساس KNN و KNN-DS ارائه دادند(۱۶). در این سیستم به منظور تشخیص حمله نوع Probe از تشخیص دو حمله نوع U2R و R2L بهره گرفته می‌شود.

وارسی(۱۷)، تحقیقی مبتنی بر الگوریتم ازدحام ذرات برای تشخیص نفوذ ارائه داد(۱۷). روش پیشنهادی یک روش ترکیبی است که بر اساس استخراج قوانین رابطه‌ای و انتخاب تکراری آنها بر اساس بهینه‌ساز ازدحام ذرات عمل می‌کند. آدینه و همکاران(۱۸)، یک سیستم تشخیص نفوذ به منظور بهبود بخش‌های ایجاد جمعیت اولیه و اپراتور انتخاب در الگوریتم ژنتیک ارائه دادند(۱۸). شده است. نتایج این تحقیق نشان می دهد که با ترکیب IDS و الگوریتم ژنتیک نرخ تشخیص نفوذ افزایش یافته است. در تحقیقی دیگر خدایار و همکاران(۱۹)، به ارائه یک روش ترکیبی یادگیری ماشین به منظور تشخیص نفوذ در شبکه می‌پردازند(۱۹). روش ترکیبی ارائه شده در این پژوهش مبتنی بر مفهوم کاهش ابعاد و الگوریتم‌های درخت تصمیم و روش‌های ترکیبی بوده و از دقت حدود ۹۷٪ برخوردار است. محمود و رائیس(۲۰)، از الگوریتم بهینه‌ساز مورچه به منظور انتخاب ویژگی در مجموعه داده‌های تشخیص نفوذ استفاده کردند(۲۰). اغلب سیستم‌های تشخیص نفوذ برای تشخیص حمله مبتنی بر ناهنجاری و یا مبتنی بر امضاء هستند. در این روش تشخیصی، کیفیت جواب به شدت به کیفیت ویژگی‌های ورودی بستگی دارد.

---

[3] Information Gain
[4] Gain Ratio

۳- روش پیشنهادی

با توجه به اینکه مجموعه داده NSL-KDDCUP مورد استفاده در این تحقیق از تعداد زیادی ویژگی برخوردار است، لذا در گام نخست با استفاده از یک الگوریتم ژنتیک(۲۱) ساده سعی در انتخاب زیر مجموعه‌ای از بهترین ویژگی‌های موثر در تشخیص نفوذ داریم. در گام بعد با توجه به ویژگی‌های انتخاب شده، یک مدل فراابتکاری جهت طبقه‌بندی داده‌ها، با استفاده از الگوریتم مورچه-ها(۲۲) و درخت تصمیم(۲۳) ارائه می‌دهیم.

به منظور آماده‌سازی داده‌ها در ابتدا پیش پردازشی روی داده‌ها انجام می‌شود. اینکار به منظور بهبود کیفیت داده‌های واقعی انجام می-شود و شامل سه مرحله «تغییر مقادیر مشخصه‌های رشته‌ای به عدد»، «در هم ریختن مجموعه داده[۵]» و «نرمال‌سازی داده‌ها» می‌باشد. با توجه به اینکه در پردازش داده‌ها با اعداد سروکار داریم و برخی از مشخصه‌های مجموعه داده مورد استفاده دارای مقادیر رشته‌ای هستند، لذا تغییر آنها به مقادیر عددی برای انجام پردازش الزامی می‌باشد. در برخی از مجموعه داده‌ها، نمونه‌های مربوط به کلاس‌های یکسان به صورت متوالی قرار گرفته‌اند، لذا برای افزایش کارایی و بالا بردن دقت روش‌های طبقه‌بندی، نمونه‌ها را به صورت تصادفی در مجموعه داده قرار می‌دهیم (تکنیک در هم ریختن). همچنین هنگامی که مقادیر مشخصه‌ها در محدوده‌های متفاوتی قرار داشته باشند، جهت کسب نتایج بهتر و کاهش پیچیدگی زمانی آنها را در دامنه‌های مشابهی قرار می‌دهیم. در این تحقیق برای نرمال‌سازی داده‌ها از روش نرمال‌سازی مینیمم – ماکزیمم مطابق با رابطه ۱ استفاده می‌شود.

$$X_j^{new} = \frac{X_j - \min_j(X_j)}{\max_j(X_j) - \min_j(X_j)} \quad (۱)$$

در این رابطه X_j مقدار ویژگی j ام برای تمام نمونه‌ها می‌باشد. مقادیر هر ویژگی در مجموعه داده به صورت مجزا نرمال‌سازی می‌شوند. در ادامه دو مرحله انتخاب ویژگی و طبقه‌بندی داده‌ها را جهت ایجاد یک سیستم تشخیص نفوذ شرح می‌دهیم.

۳-۱- انتخاب ویژگی‌ها مبتنی بر الگوریتم ژنتیک

الگوریتم‌های ژنتیک یکی از الگوریتم‌های جستجوی تصادفی است که ایده آن برگرفته از طبیعت می‌باشد. هدف این بخش سفارشی کردن الگوریتم ژنتیک به منظور انتخاب بهترین زیر مجموعه از ویژگی‌ها به صورت خودکار می‌باشد. در این تحقیق برای مسئله انتخاب ویژگی‌ها، شماره هر ویژگی را به عنوان معیار انتخاب آن ویژگی در نظر می‌گیریم. در این صورت ویژگی‌هایی که شماره آن در کروموزم وجود داشته باشد، جزء ویژگی‌های مطلوب[۶] خواهند بود. در این حالت طول کروموزم، تعداد ویژگی‌های مطلوب خواهد بود.

الگوریتم ژنتیک نیز کار خود را با تولید جمعیت اولیه از کروموزوم‌ها به صورت تصادفی از فضای جستجو آغاز می‌کند. نکته مهم هنگام تولید جمعیت اولیه مشخص کردن طول کروموزوم است. در اغلب روش‌های انتخاب ویژگی، تعداد ویژگی‌های مطلوب به عنوان ورودی به الگوریتم داده می‌شود. این تعداد با توجه به مسئله و ابعاد مجموعه داده به صورت دستی و با سعی و خطا بدست می‌آید. در مجموعه داده‌های ابعاد بالا نظیر NSL-KDDCUP، یافتن تعداد ویژگی‌های مطلوب سخت و نیاز به انجام آزمایش‌های زیادی دارد. لذا در این تحقیق رویکردی را پیشنهاد می‌دهیم که تعداد ویژگی‌های مطلوب به صورت خودکار مشخص شود.

در اینجا ساختار جمعیت اولیه را به صورت یک سلول با NP (اندازه جمعیت) کروموزوم در نظر می‌گیریم. هر کروموزوم از این سلول یک راه‌حل با طولی در محدوده ۱ تا NF (تعداد کل ویژگی‌ها) می‌باشد. نکته قابل توجه این است که باید با تمامی تعداد

[5] Dataset Shuffle
[6] Desired Number Of Features

ویژگی‌ها در کروموزوم وجود داشته باشد. به عبارت دیگر باید جمعیت ایجاد شده تمامی تعداد ویژگی‌های مطلوب را پوشش دهد. این کار مستلزم آن است که شرط NP>NF همیشه برقرار باشد. این شرط این اطمینان را می‌دهد که از تمامی‌طول ها حداقل یک کروموزوم وجود خواهد داشت. برای اینکه توزیع تمامی کروموزوم‌ها با طول‌های متفاوت در جمعیت برابر باشند، از هر کروموزوم با یک طول مشخص $NP/_{NF}$ کروموزوم ساخته می‌شود. نمونه‌ای از جمعیت کروموزوم‌های پیشنهادی را در شکل ۱ مشاهده می‌کنید.

کروموزوم با طول ۱	۸		
کروموزوم با طول ۲	۵	۲	
کروموزوم با طول ۳	۹	۵	٤
...	–	–	–
کروموزوم با طول NF	–	–	–
کروموزوم با طول ۱	٦		
کروموزوم با طول ۲	۷	۳	
...	–	–	–

شکل ۱ : نمونه‌ای از جمعیت کروموزوم‌های پیشنهادی

در این تحقیق کروموزوم‌ها با استفاده از درخت تصمیم(DT) مورد ارزیابی قرار می‌گیرند و از روش تورنومنت[7] برای انتخاب والدین جهت تولید و مثل استفاده می‌شود. عملگر ترکیب ارائه شده در این تحقیق به احتمال CR از بین ویژگی‌های دو کروموزوم والد، ۸۰ درصد از ویژگی‌های کروموزوم فرزند را به صورت تصادفی انتخاب می‌کند. سایر ویژگی‌ها به صورت تصادفی از بین ویژگی‌هایی انتخاب می‌شود که در دو کروموزوم والد وجود نداشته باشند. همانطور که گفته شد هر کروموزوم می‌تواند با طول متفاوتی از فضای جستجو ایجاد شود. لذا طول کروموزوم فرزند به صورت تصادفی از بین طول دو کروموزوم والد انتخاب می‌شود. بعد از تولید فرزند، عملگر جهش به احتمال MR روی هر ژن از کروموزوم تولید شده اعمال شده و محتوای آن را به صورت تصادفی تغییر می‌دهد. در نهایت فرزند تولید شده بعد از این مرحله مورد ارزیابی قرار گرفته و در صورت بهبود دقت نسبت به والد متناظر خود، جایگزین آن می‌شود.

با توجه به اینکه هر کروموزوم دارای احتمال خاصی برای شرکت در عملگرهای ترکیب و جهش می‌باشد، از طریق کنترل تطبیقی[8] بهبود فوق العاده‌ای در توانایی جستجو الگوریتم ژنتیک حاصل می‌شود(۲۱). دو احتمال Cr و Mr تا حد زیادی میزان دقت راه حل ها و سرعت همگرایی الگوریتم ژنتیک را تعیین می کند که از طریق رابطه تطبیقی ۲ و ۳ محاسبه می‌شود.

$$Cr = \max(\alpha, Cr \times \beta) \quad (۲)$$

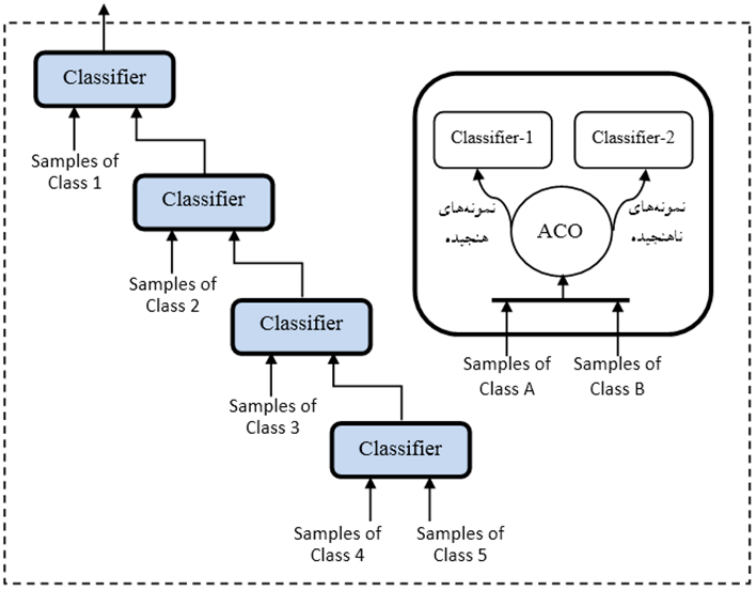
$$Mr = \max(\alpha, Mr \times \beta) \quad (۳)$$

در این روابط α و β دو مقدار ثابت کوچکتر از یک هستند که سرعت کاهش Cr و Mr را کنترل می‌کنند.

[7] Tournament
[8] Adaptive

۳-۲- طبقه‌بندی داده‌ها مبتنی بر الگوریتم مورچگان و روش یادگیری زوجی

جهت تشخیص رکوردهای نفوذ در شبکه یک روش طبقه‌بندی فراابتکاری بر مبنای یادگیری زوجی و الگوریتم کلونی مورچگان ارائه می‌دهیم. شکل ۲ فلوچارت روش طبقه‌بندی پیشنهادی را نشان می‌دهد.



شکل ۲ : فلوچارت روش طبقه‌بندی پیشنهادی مبتنی بر الگوریتم مورچگان و یادگیری زوجی

با توجه به ۵ کلاسه بودن مجموعه داده مورد ارزیابی، در روش زوجی ارائه شده باید به ازای هر جفت کلاس یک مدل طبقه‌بندی در مرحله آموزش ایجاد شود. در مرحله آزمایش نیز، نمونه جدید با توجه به ٤ طبقه‌بند متمایز مطابق شکل ۲ کلاس خروجی را پیش‌بینی می‌کند. ورودی طبقه‌بند اول، تنها نمونه‌هایی از مجموعه داده با کلاس‌های ٤ و ۵ می‌باشد. خروجی این طبقه‌بند میزان تشابه نمونه ورودی را به دو کلاس ٤ و ۵ مشخص می‌کند که به عنوان ورودی طبقه‌بند بعدی در نظر گرفته می‌شود. در طبقه‌بند دوم باید مدل‌های طبقه‌بند با توجه به کلاس خروجی نمونه‌ها در طبقه‌بند قبل، بکار گرفته شوند. بنابراین نمونه‌های خروجی از طبقه‌بند اول که به کلاس ٤ یا ۵ شبیه هستند، میزان تشابه آنها به کلاس ۳ نیز بررسی می‌شود. این سناریو تا پایان طبقه‌بند دوم، میزان تشابه نمونه ورودی را بین سه کلاس ۵، ٤ و ۳ مشخص کرده و کلاس با بالاترین تشابه را معرفی می‌کند. این روند برای طبقه‌بندهای سوم و چهارم نیز به همین صورت اعمال می‌شود و در نهایت کلاس نمونه ورودی تشخیص داده می‌شود. طبقه‌بند استفاده شده در این تحقیق درخت تصمیم می‌باشد.

با توجه به استاندارد بودن مدل طبقه‌بندی درخت تصمیم و به منظور افزایش دقت آن یک روش فراابتکاری برای مدل طبقه‌بندی ارائه می‌دهیم. فرض کنید یک مجموعه داده با ۱۰۰ نمونه و n کلاس موجود باشد. طبقه‌بند درخت تصمیم این مجموعه داده را در حالت آموزش با دقت ۹۵٪ مدل می‌کند. مشخص است که ۵ نمونه از این مجموعه داده به اشتباه کلاس‌بندی شده‌اند. حال اگر مجموعه داده اولیه را بدون این ۵ نمونه مجددا طبقه‌بندی کنیم، انتظار می‌رود دقت حاصل شده افزایش یابد. برخی از نمونه‌های بعضی مجموعه داده‌ها، ممکن است در حالات بسیار نادر و خاصی با مقادیر ویژگی‌های متفاوتی ایجاد شده باشند که باعث تمایز آنها از سایر نمونه‌ها شده است. با تشخیص و جدا کردن این نمونه‌ها از مجموعه داده می‌توان دقت مدل آموزشی را افزایش داد. در این تحقیق به نمونه-هایی با این خصوصات «ناهنجیده» و به سایر نمونه‌ها «هنجیده» می‌گوییم.

در مدل طبقه‌بندی ارائه شده برای هر زوج کلاس ابتدا کل مجموعه داده را به دو بخش «هنجیده» و «ناهنجیده» تقسیم می‌کنیم. هدف

از این تقسیم‌بندی مشخص کردن، کمترین تعداد نمونه‌هایی می‌باشد که با جداسازی آنها از مجموعه داده، دقت آموزش حداکثر شود. برای تحقق این هدف از الگوریتم بهینه‌سازی مورچه‌ها و خصوصیت آن استفاده می‌کنیم. در مرحله بعد نمونه‌های بخش «هنجیده» و «ناهنجیده» به صورت مجزا طبقه‌بندی می‌شوند. بنابراین برای یک نمونه ورودی، ابتدا نوع آن مشخص می‌شود(«هنجیده» یا «ناهنجیده»)، سپس با توجه به نوع، طبقه‌بند مربوطه کلاس خروجی نمونه را تعیین می‌کند. در ادامه روند جداسازی نمونه‌ها به دو بخش «هنجیده» و «ناهنجیده» با استفاده از الگوریتم بهینه‌سازی مورچه‌ها شرح داده می‌شود.

استفاده از الگوریتم مورچگان نخستین بار توسط دوریگو و همکاران(٢٤)، در سال ۱۹۹۱ برای حل مسائل بهینه‌سازی پیچیده از جمله فروشنده دوره گرد(TSP) ارائه گردید و مورد توجه محققان قرار گرفت(٢٤). هدف از ارائه الگوریتم مورچه‌ها در اینجا، تشخیص نمونه‌های «هنجیده» و «ناهنجیده» می‌باشد. نمایش مورچه‌ها در فضای جستجو، به صورت یک آرایه تک بعدی به طول کل نمونه‌ها در نظر گرفته می‌شود. هر درایه، مشخصه نمونه متناظر با آن در راه‌حل است. ارزش هر اندیس به صورت باینری در نظر گرفته شده است، 0 به معنی «ناهنجیده» بودن و 1 به معنی «هنجیده» بودن آن نمونه در راه‌حل است. در این تحقیق از یک الگوریتم حریصانه به منظور تولید مورچه‌ها از فضای جستجو استفاده می‌کنیم. برای ایجاد مورچه‌ها ابتدا با استفاده از مدل طبقه‌بندی درخت تصمیم نمونه-های دو کلاس زوجی ورودی را به دو دسته ۱ و ۲ طبقه‌بندی می‌کنیم. سپس برای مورچه اول، برای نمونه‌هایی با طبقه‌بندی درست، در اندیس‌های متناظر با آنها عدد 1 و برای نمونه‌هایی که به اشتباه طبقه‌بندی شده‌اند، عدد 0 را قرار می‌دهیم. برای مورچه دوم، بدون درنظر گرفتن نمونه‌هایی با طبقه‌بندی اشتباه از مرحله قبل، طبقه‌بندی داده‌ها را مجددا تکرار می‌کنیم. اکنون نمونه‌هایی که درست طبقه‌بندی شده‌اند را برای مورچه دوم در اندیس‌های متناظر 1 و برای سایر نمونه‌ها (نمونه‌هایی که به اشتباه در این مرحله و مرحله قبل طبقه‌بندی شده‌اند) 0 را قرار می‌دهیم. این مراحل تا زمانی ادامه پیدا می‌کند که همه مورچه‌ها مقداردهی شده باشند یا اینکه در یک مرحله از طبقه‌بندی هیچ نمونه‌ای به اشتباه طبقه‌بندی نشده باشد. شبه کد الگوریتم حریصانه پیشنهادی در شکل ۳ نشان داده شده است. بعد از قرار دادن مورچه‌ها در فضای جستجو باید مقدار شایستگی هر مورچه محاسبه شود. نمونه‌ها برای هر مورچه به دو دسته تقسیم شده‌اند، لذا نیازمند اعمال دو طبقه‌بند مختلف می‌باشد. همچنین یک طبقه‌بند ابتدایی در هر مورچه برای تشخیص نمونه‌های «هنجیده» و «ناهنجیده» لحاظ می‌شود. در نهایت برای محاسبه شایستگی مورچه k ام (η^k) از رابطه ٤ استفاده می‌کنیم.

$$\eta^k = \frac{\text{classify1} + \text{classify2} + \text{classify3}}{3} \qquad (٤)$$

در این رابطه classify1 مدل طبقه‌بندی برای تشخیص نمونه‌های «هنجیده» و «ناهنجیده»، classify2 مدل طبقه‌بندی برای نمونه‌های «هنجیده» و classify3 مدل طبقه‌بندی برای نمونه‌های «ناهنجیده» می‌باشد.

Greedy Algorithm To Initialize The Ants
Input : samples of datasets for two couple class

newData = Change Samples Class to 1 and 2
For i = 1 to nAnts **do**
DT = Decision tree model(newData)
Ants(i,DT.Correct) = 1 and Ants(i,Other Samples) = 0
If (DT.Incorrect == Null)
Break
End If
newData = Remove samples that have been incorrectly classified of newData
End For

Output : Ants

شکل ۳ : شبه کد الگوریتم حریصانه جهت مقداردهی اولیه مورچه‌ها

برای انجام الگوریتم کلونی مورچگان نیازمند یک حافظه مشترک بین مسیرها می‌باشد. در این تحقیق حافظه مشترک با ایجاد یک ماتریس فرومون روی نمونه‌ها حاصل می‌شود. بنابراین یک ماتریس یک بعدی به طول تعداد نمونه‌ها با مقداردهی اولیه تصادفی در بازه $[0,1]$ ایجاد می‌کنیم. به منظور حرکت مورچه‌ها ابتدا باید مسیر حرکت هر مورچه مشخص شود. انتخاب مسیر با استفاده از یک تابع احتمالی به فرم رابطه ۵ محاسبه می‌شود.

$$P_l^k = \frac{\tau_i^\alpha \times \eta_k^\beta}{\sum_l \tau_i^\alpha \times \eta_k^\beta} \qquad (۵)$$

در این رابطه τ_i نماد مقدار فرمون نمونه i می‌باشد($i = 1, 2, 3, \ldots, l$). η_k شایستگی مورچه k ام و α و β دو پارامتر ثابت در بازه $[0,1]$ هستند که تاثیر نسبی بین فرومون و شایستگی را تنظیم می‌کند.

نکته حائز اهمیت در اینجا مقدار l است که مسیرهای قابل حرکت برای مورچه را نشان می‌دهد. با توجه به خصوصیت مسئله مورچه توانایی تغییر هر نمونه‌ای را از 0 به 1 و بلعکس دارد. این موضوع باعث افزایش ابعاد مسئله و کاهش سرعت همگرایی مسئله خواهد شد. به منظور کاهش ابعاد حرکتی هر مورچه، تنها از نمونه‌هایی که باعث خطا در مدل طبقه‌بندی ابتدایی می‌شوند، استفاده می‌کنیم. به این معنا که مورچه فقط قادر به تغییر نمونه‌هایی است که در classify1 به اشتباه طبقه‌بندی شده‌اند.

با محاسبه احتمال حرکت هر مورچه و با استفاده از سیاست انتخاب تورنومنت یکی از مسیرها انتخاب و مورچه به سمت آن حرکت می‌کند. حرکت مورچه به سمت مسیر انتخابی با تغییر بیت نمونه مورد نظر انجام می‌شود. با توجه به تعداد زیاد نمونه‌ها، در هر مرحله، φ نمونه در یک مورچه تغییر داده می‌شود. بعد از حرکت هر مورچه، میزان شایستگی مورچه در مکان جدید محاسبه شده و در صورت بهبود دقت نسبت به مکان قبلی، مشخصات مورچه‌ها با توجه به مکان جدید به روز می‌شوند.

بعد از این که هر مورچه مقصد بعدی خود را انتخاب کرد، به وسیله رابطه ٦ فرومون آن مسیر(نمونه) به روزرسانی می‌گردد. این به روزرسانی محلی می‌باشد، چون تنها روی مسیری که مورچه حرکت کرده است انجام می‌شود.

$$\tau_i^k = (1-\rho)\tau_i^k + \rho\Delta\tau_i^k \quad , \qquad \Delta = {}^{1}/_{(100 - \eta_k)} \qquad (٦)$$

در این رابطه ρ ضریبی بین $[0,1]$ است که تبخیر فرومون هر مسیر را مشخص می‌کند. τ_i^k میزان فرمون جاری روی نمونه i ام برای مورچه k ام است. Δ میزان افزایش فرمون روی مسیر جاری که با توجه به خطای طبقه‌بندی مشخص می‌شود.

به روز کردن سراسری فرومون هنگامی که تمامی مورچه‌ها یک دور کامل را پشت سر گذاشته باشند توسط مورچه‌ای که بهترین مسیر را طی کرده است انجام می‌گیرد. رابطه ۷ برای بروزرسانی سراسری استفاده می‌شود.

$$\tau_i^{best} = (1-\rho)\tau_i^{best} + \rho\Delta\tau_i^{best} \quad (۷)$$

در این رابطه τ_i^{best} اطلاعات فرمون بهترین مورچه می‌باشد. شبه کد الگوریتم طبقه‌بندی پیشنهادی به منظور تشخیص نمونه‌های «هنجیده» و «ناهنجیده» در شکل ٤ آورده شده است.

Ant Colony Optimization Algorithm For Classification
Input : Samples of datasets with features selected

Place each ant on initial node (Greedy Algorithm)
Set pheromone on the samples to small constant
Calculation of the merits of each ant, (eq. 4)
For i=1 to iteration **do**
  For k=1 to nAnts **do**
  Calculates directions probability of k ant, (eq. 5)
  Move ant towards the path with φ dimensions
  Apply local update pheromone, (eq. 6)
  End For
  Apply global update pheromone, (eq. 7)
End For

Output : Ant Best

شکل ٤ : شبه کد الگوریتم طبقه‌بندی پیشنهادی

٤- نتایج و آزمایشات

در این تحقیق از مجموعه داده NSL-KDDCUP به منظور ارزیابی عملکرد روش پیشنهادی استفاده شده است(۲۵). این مجموعه داده دارای ۵ کلاس Normal، U2R، R2L، Dos و Probe می‌باشد. همچنین به منظور پیاده‌سازی مدل ارائه شده از نرم افزار متلب ورژن ۲۰۱٦ شده است. نتایج بدست آمده از این بررسی به منظور حصول نتایج دقیق تر، میانگین ۲۵ مرتبه تکرار تست می‌باشد. مقادیر پارامترهای استفاده شده در شبیه‌سازی در جدول ۱ نشان داده شده است.

جدول ۱ : پارامترهای استفاده شده در شبیه‌سازی

پارامترهای الگوریتم ژنتیک			**پارامترهای الگوریتم مورچگان**		
نام پارامتر	**مقدار**	**توضیحات**	**نام پارامتر**	**مقدار**	**توضیحات**
NP	50	اندازه جمعیت	nAnt	35	تعداد مورچه
CR	0.85	نرخ ترکیب	iteration	75	تعداد تکرار
MR	0.20	نرخ جهش	φ	5	گام حرکت
Gen	30	تعداد نسل	ρ	0.3	نرخ تبخیر فرومون
(α,β)	(0.1,0.95)	کنترل تطبیقی	(α,β)	(0.5,0.75)	ضریب احتمال

ارزیابی یک مدل طبقه‌بندی بر اساس نمونه‌های آموزشی و آزمایشی صورت می‌گیرد. وقوع حالات مختلف برای دسته‌ها برای طبقه-بندی، با مقادیر FN، TN، FP و TP برای یک سیستم تشخیص نفوذ در حالت دو کلاسه قابل نمایش است(۲۶). مهمترین معیاری که برای تعیین کارایی یک الگوریتم طبقه‌بندی استفاده می‌شود، معیار دقت است که نشان می‌دهد چند درصد از کل مجموعه داده به درستی تشخیص داده شده‌اند. رابطه ۸ نحوه محاسبه این معیار را نشان می‌دهد.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (۸)$$

دو مقدار TN و TP از مهمترین مقادیری هستند که برای به حداکثر رسیدن کارایی طبقه‌بندی باید بیشینه شوند. معیار دقت در مجموعه داده‌هایی با دسته‌های نامتعادل و تعداد نمونه‌های مختلف، معیار مناسبی نمی‌باشد. لذا نیاز به معیارهای مناسب نظیر precision و recall است. معیار precision دقت طبقه‌بندی دسته i را با توجه به کل مواردی نشان می‌دهد که برچسب i برای نمونه مورد بررسی توسط طبقه‌بند پیشنهاد شده است. نحوه محاسبه این معیار در رابطه ۹ نشان داده شده است.

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (۹)$$

معیار recall دقت طبقه‌بندی دسته i را با توجه به کل نمونه‌های متعلق به برچسب i نشان می‌دهد که با توجه به رابطه ۱۰ محاسبه می‌شود.

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (۱۰)$$

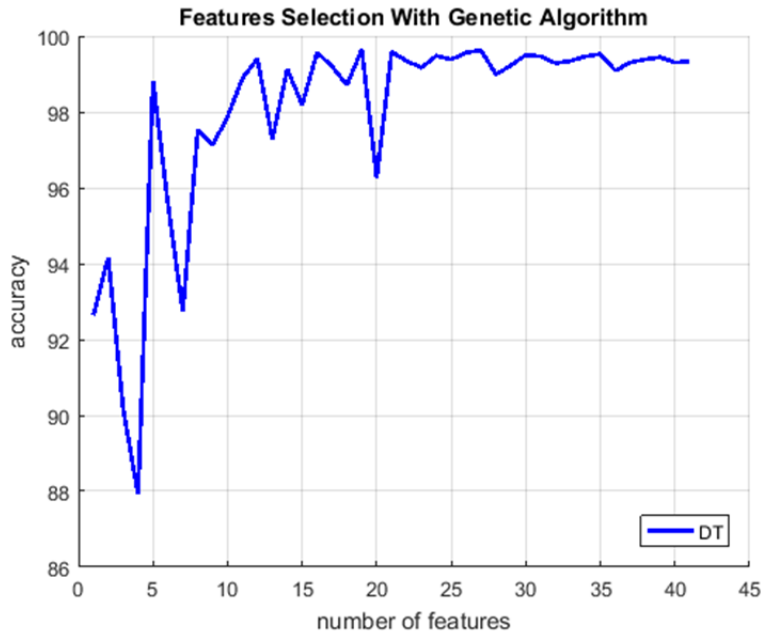
معیار f-measure از ترکیب معیارهای precision و recall بدست می‌آید و به صورت رابطه ۱۱ می‌باشد. این معیار در مواردی که نتوان اهمیت ویژه‌ای را برای هر یک از دو معیار precision و recall نسبت به یکدیگر قائل شد، مورد استفاده قرار می‌گیرد.

$$F - measure = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (۱۱)$$

با توجه به خاصیت طول رشته متغیر در الگوریتم پیشنهادی، برای نشان دادن ویژگی‌های انتخاب شده میانگین فراوانی ویژگی‌ها از اجراهای مختلف در نظر گرفته می‌شود. ویژگی‌های انتخاب شده توسط الگوریتم ژنتیک در جدول ۲ نشان داده شده است. تعداد ۱۹ ویژگی از ۴۱ ویژگی موجود در مجموعه داده انتخاب شده است. دقت الگوریتم ژنتیک در انتخاب ویژگی با تعداد متفاوت را شکل ۵ مشاهده می‌کنید.

جدول ۲ : ویژگی‌های انتخاب شده توسط الگوریتم ژنتیک

No.	attribute name	No.	attribute name	No.	attribute name
1	duration	9	urgent	25	serror_rate
3	service	10	hot	26	srv_serror_rate
4	flag	11	num_failed_logins	33	dst_host_srv_count
5	src_bytes	12	logged_in	38	dst_host_serror_rate
6	dst_bytes	15	su_attempted	41	dst_host_srv_rerror_rate
7	land	16	num_root	-	-
8	wrong_fragment	23	count	-	-



شکل ٥ : دقت الگوریتم ژنتیک در انتخاب ویژگی با تعداد متفاوت

ماتریس درهم ریختگی داده‌های سیستم تشخیص نفوذ به تفکیک دسته‌ها در جدول ۳ آورده شده است. در این جدول تعداد رکوردها برای هر نوع حمله به همراه تعداد تشخیص‌های درست ذکر شده است.

جدول ۳ : ماتریس درهم ریختگی به تفکیک نوع حمله

Actual Records			Number of predicted				
Records Type	**Dataset**	**Number**	**Normal**	**DOS**	**U2R**	**R2L**	**Probe**
Normal	Train	67343	67338	0	0	0	5
	Test	9710	9705	1	0	0	4
DOS	Train	45927	0	45919	0	0	8
	Test	7458	28	7430	0	0	0
U2R	Train	52	1	1	50	0	0
	Test	200	2	0	198	0	0
R2L	Train	995	1	1	0	993	0
	Test	2754	7	0	0	2747	0
Probe	Train	11656	3	1	0	0	11652
	Test	2421	11	0	0	0	2410

در ادامه برای ارزیابی هر چه بیشتر رویکرد فوق، عملکرد سیستم پیشنهادی را با سایر روش‌های تشخیص نفوذ مشابه که نتایج آزمایشات خود را بر روی داده‌های NSL-KDDCUP ارائه کرده‌اند، مقایسه می‌کنیم. نتایج حاصل از مقایسه در جدول ٤ نشان داده شده است. روش تشخیص نفوذ پیشنهادی در مقایسه با سایر روش‌های تشخیص نفوذ و به ازای برخی از حملات دقت بیشتری داشته و در بقیه موارد نیز دقت مناسبی را ارائه می‌دهد. در جدول ارائه شده مقادیر هر یک از کلاس‌ها بر اساس مقادیر ارائه شده در تحقیقات مربوطه می‌باشد لذا برخی از فیلدها ممکن است در تحقیقات ذکر نشده باشند.

جدول ٤ : مقایسه روش‌های تشخیص نفوذ به تفکیک نوع حمله با معیارهای مختلف (٪)

Methods	Category	Normal	DOS	U2R	R2L	Probe	All
Fuzzy + ACO (15)	accuracy	-	-	-	-	-	99.69
ACO + SVM (20)	accuracy	-	-	-	-	-	98.29
IDS ACO (11)	accuracy	97.41	99.78	93.51	99.17	74.65	98.9
	precision	61.31	79.22	8.17	93.14	85.79	-
	recall	97.65	75.26	73.55	24.66	68.86	-
	f-measure	75.33	77.19	14.71	39.00	76.40	-
FARCHD-OVO (27)	accuracy	99.81	98.05	65.38	87.54	95.83	99.00
	precision	98.63	99.84	95.28	23.29	97.87	-
	recall	99.81	98.05	65.38	87.54	95.83	-
SIPSO (17)	accuracy	-	99.80	97.50	82.50	99.70	-
CSM (9)	accuracy	-	-	-	-	-	99.79
RandomTree (19)	accuracy	**99.85**	**100.0**	88.45	98.00	**99.70**	-
MARS	accuracy	99.71	99.97	76.00	98.75	99.85	–
Proposed Method	accuracy	99.81	99.74	**99.39**	**99.76**	99.68	**99.79**
	precision	100	99.61	99.00	99.74	99.54	-
	recall	99.62	99.86	99.79	99.79	99.81	-
	f-measure	99.81	99.74	99.29	99.76	99.68	-

یکی از دلایل عملکرد خوب روش پیشنهادی استفاده از الگوریتم مورچگان برای تشخیص و جداسازی نمونه‌های متفاوت با کلاس مربوطه می‌باشد. این نمونه‌ها ممکن است به ندرت اتفاق بیافتد و یا نویز باشند. بنابراین روش پیشنهادی استراتژی مقابله با نویز را نیز خواهد داشت. این مورد در دسته U2R که دارای نمونه‌های آموزشی کمتری است به خوبی مشهود است و الگوریتم پیشنهادی در این دسته نتایج خوبی را به ثبت رسانیده است.

۵- نتیجه‌گیری و پیشنهادات

هدف نهایی سیستم‌های تشخیص نفوذ به کارگیری الگوریتم‌های داده کاوی به منظور تشخیص صحیح و برخط نفوذها می‌باشد. دقت الگوریتم‌های داده کاوی بستگی به انتخاب ویژگی‌های مناسب و همچنین تعداد رکوردهای مورد نظر برای یادگیری است. الگوریتم ژنتیک ارائه شده به خوبی ویژگی‌های مناسب را انتخاب کرده و همچنین ساختار طول رشته متغیر آن باعث شده تعداد ویژگی‌ها به صورت خودکار تعیین شود. روش طبقه‌بندی ترکیبی ارائه شده نیز به خوبی دسته‌بندی داده‌ها را انجام می‌دهد. این روش مبتنی بر تشخیص نمونه‌های متناقض می‌باشد که به وسیله الگوریتم مورچگان تشخیص داده می‌شود. الگوریتم مورچه‌ها در هر بخش وظیفه تشخیص نمونه‌های متناقض (شبیه به نویز) را بر عهده دارد. با بخش‌بندی نمونه‌های ورودی به دو بخش متناقض و معتبر، مدل طبقه‌بندی روی هر دو بخش ایجاد می‌شود. نتایج حاصل از آزمایشات کارایی بهتر الگوریتم پیشنهادی را خصوصا در دسته‌هایی با نمونه‌های کمتر نشان می‌دهد.

با توجه به اینکه ترتیب ورودی نمونه‌ها به هر مورچه در این تحقیق لحاظ نشده است، برای تحقیقات آتی پیشنهادی می‌شود مدلی با این قابلیت طرحی و عملکرد آن مورد ارزیابی قرار گیرد. برای مثال اضافه کردن یک الگوریتم جستجو محلی برای پیدا کردن ترتیب مناسبی از نمونه‌ها قبل از ایجاد جمعیت اولیه به سیستم می‌تواند در افزایش دقت سیستم تشخیصی موثر باشد.

منابع

(1) Debar, H. (2000). An introduction to intrusion-detection systems. Proceedings of Connect, 2000.
(2) Mukherjee, B., Heberlein, L. T., & Levitt, K. N. (1994). Network intrusion detection. IEEE network, 8(3), 26-41.
(3) Raghunath, B. R., & Mahadeo, S. N. (2008, July). Network intrusion detection system (NIDS). In Emerging Trends in Engineering and Technology, 2008. ICETET'08. First International Conference on (pp. 1272-1277). IEEE.
(4) Mazzariello, C., Bifulco, R., & Canonico, R. (2010, August). Integrating a network ids into an open source cloud computing environment. In Information Assurance and Security (IAS), 2010 Sixth International Conference on (pp. 265-270). IEEE.
(5) Modi, C., Patel, D., Borisaniya, B., Patel, H., Patel, A., & Rajarajan, M. (2013). A survey of intrusion detection techniques in cloud. Journal of Network and Computer Applications, 36(1), 42-57.
(6) Goyal, Anup, and Chetan Kumar, (2008). GA-NIDS: a genetic algorithm based network intrusion detection system. Northwestern university, 178(15), 3024–3042
(7) Muda, Z., Yassin, W., Sulaiman, M. N., & Udzir, N. I. (2011, July). Intrusion detection based on K-Means clustering and Naïve Bayes classification. In Information Technology in Asia (CITA 11), 2011 7th International Conference on (pp. 1-6). IEEE.
(8) Saha, S., Sairam, A. S., Yadav, A., & Ekbal, A. (2012, August). Genetic algorithm combined with support vector machine for building an intrusion detection system. In Proceedings of the International Conference on Advances in Computing, Communications and Informatics (pp. 566-572). ACM.
(9) Chae, Hee-su, Byung-oh Jo, Sang-Hyun Choi, and Twaekyung Park, (2015). Feature Selection for Intrusion Detection using NSL-KDD. Recent Advances in Computer Science, ISBN : 978-960.
(10) Benaicha, S. E., Saoudi, L., Guermeche, S. E. B., & Lounis, O. (2014, August). Intrusion detection system using genetic algorithm. In Science and Information Conference (SAI), 2014 (pp. 564-568). IEEE.
(11) Aghdam, Mehdi Hosseinzadeh, and Peyman Kabiri, (2016). Feature selection for intrusion detection system using ant colony optimization. International Journal of Network Security 18.3 : 420-432.
(12) Mubarak, Shaik Liyakhat, (2016). Intrusion Detection System using SVM, SOM & NN.
(13) Kevric, J., Jukic, S., & Subasi, A. (2016). An effective combining classifier approach using tree algorithms for network intrusion detection. Neural Computing and Applications, 1-8.
(14) Chen, M. H., Chang, P. C., & Wu, J. L. (2016). A population-based incremental learning approach with artificial immune system for network intrusion detection. Engineering Applications of Artificial Intelligence, 51, 171-181.
(15) Varma, P. R. K., Kumari, V. V., & Kumar, S. S. (2016). Feature Selection Using Relative Fuzzy Entropy and Ant Colony Optimization Applied to Real-time Intrusion Detection System. Procedia Computer Science, 85, 503-510.
(16) Rawat, A., & Choubey, A. (2016). Ant Colony Optimization for Intrusion Detection System Based on KNN and KNN-DS with detection of U2R, R2L attack for Network Probe Attack Detection.
(17) Warsi, Sana, Yogesh Rai, and Santosh Kushwaha. "Selective Iteration based Particle Swarm Optimization (SIPSO) for Intrusion Detection System." International Journal of Computer Applications 124.17 (2015).
(18) Salah Eddine, Benaicha, et al. "Intrusion detection system using genetic algorithm." Science and Information Conference (SAI), 2014. IEEE, 2014.
(۱۹) خدایار, محمد؛ علیرضا عصاره و منصور امینی لاری، ۱۳۹۳، بکارگیری الگوریتم های ترکیبی یادگیری ماشین در بهبود سیستم

های تشخیص نفوذ، همایش ملی مهندسی رایانه و مدیریت فناوری اطلاعات، تهران، شرکت علم و صنعت طلوع فرزین.
(20) Mehmod, T., & Rais, H. B. M. (2016). Ant Colony Optimization and Feature Selection for Intrusion Detection. In Advances in Machine Learning and Signal Processing (pp. 305-312). Springer International Publishing.
(21) Galletly, J. E. (1992). An overview of genetic algorithms. Kybernetes, 21(6), 26-30.
(22) Maniezzo, V., & Carbonaro, A. (2002). Ant colony optimization: an overview. In Essays and surveys in metaheuristics (pp. 469-492). Springer US.
(23) Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), 660-674.
(24) Dorigo, M. (1991). Ant Colony Optimization—new optimization techniques in engineering. Springer-Verlag, Berlin Heidelberg, 101-117.
(25) Nsl-kdd data set for network based intrusion detection systems. Available on: http://nsl.cs.unb.ca/KDD/NSL-KDD.html, March 2009.
(26) Bates, D. W., Goldman, L., & Lee, T. H. (1991). Contaminant blood cultures and resource utilization: the true consequences of false-positive results. JAMA, 265(3), 365-369.
(27) Elhag, S., Fernández, A., Bawakid, A., Alshomrani, S., & Herrera, F. (2015). On the combination of genetic fuzzy systems and pairwise learning for improving detection rates on Intrusion Detection Systems. Expert Systems with Applications, 42(1), 193-202.

Using Ant Colony Algorithm and Pairwise Learning to Classify Attack in Intrusion Detection Systems

Mohammad Ali Nadoomi[1], Sina Majid[2]

Computer engineering,Advanced Studies ,Islamic Azad university,Boushehr, Iran, nadoomi@gmail.com
Islamic Azad university, Boushehr, majidsina.edu@gmail.com

Abstract

Intrusion detection systems for security in computer networks have been proposed to be crossed if the attacker from other security equipment, able to detect it and prevent it from advancing. One of the challenges of these systems, it is high dimensional data. In this study was to reduce the dimensions of a simple genetic algorithm with the length of the string variable we use. Then, according to selected characteristics, a meta-heuristic model for data classification, using ant colony algorithm offer. Classification model proposed by trying to divide the data into two samples is Hnjydh and Nahnjydh. The proposed method for evaluating the performance of database intrusion detection NSL-KDD than other data from the records of more realistic approach is used. The results of the experiments, the proposed method has better performance compared with other existing methods show.

Keywords: Feature selection, classification, genetic algorithm, ant colony algorithm, database NSL-KDD.