تشخیص نفوذ و ارتقا امنیت در شبکه صداوسیمای مرکز بوشهر با استفاده از PCA و طبقه‌بندی ترکیبی

مریم جوکار[۱]، موسی مجرد[۲،۳] و حسن ارفعی‌نیا[۴]

[۱]دانشجو کارشناسی ارشد، موسسه آموزش عالی غیرانتفاعی-غیر دولتی لیان بوشهر، بوشهر، ایران، mar.mj324@gmail.com

[۲]استادیار، دانشکده مهندسی برق و کامپیوتر، واحد فیروزآباد، دانشگاه آزاد اسلامی، فیروزآباد، ایران، m.mojarad@iauf.ac.ir

[۳]باشگاه پژوهشگران جوان، واحد فیروزآباد، دانشگاه آزاد اسلامی، فیروزآباد، ایران

[۴]مربی، موسسه آموزش عالی غیرانتفاعی-غیر دولتی لیان بوشهر، بوشهر، ایران، harfaeinia@gmail.com

چکیده: با توجه به اهمیت پرداختن به میزان اثربخشی برنامه‌های صداوسیما و نیز لزوم اعتمادسازی به آن، بحث حملات سایبری و نفوذ آن چندان مورد توجه قرار نگرفته است. همین امر امنیت شبکه‌های وابسته به این سیستم‌ها را با چالشی بزرگ روبرو کرده است. این در حالی است که سرعت بالای توسعه فن‌آوری، اهمیت توجه به مسئله تشخیص نفوذ در این حوزه را دوچندان کرده است. هدف اصلی پژوهش حاضر شناسایی ویژگی‌های مؤثر در شبکه‌های وابسته به صداوسیمای مرکز بوشهر است که در آن پژوهشگر به دنبال چگونگی مدل‌سازی داده‌ها جهت تشخیص حملات سایبری و ورودهای غیرمجاز است. در این تحقیق تشخیص نفوذ و ارتقا امنیت در شبکه صداوسیمای مرکز بوشهر با استفاده از الگوریتم PCA و یک روش طبقه‌بندی ترکیبی انجام شده است. در اینجا الگوریتم PCA برای کار انتخاب ویژگی‌های مؤثر و روش طبقه‌بندی ترکیبی برای کار مدل‌سازی داده‌ها و تشخیص نفوذ استفاده می‌شود. دلیل استفاده از مدل طبقه‌بندی ترکیبی توانایی بیشتر آن در پیش‌بینی داده‌ها و مدل‌سازی الگوهای پیچیده‌تر است. پژوهش از نظر هدف کاربردی و به روش توصیفی انجام شده است. جامعه آماری تحقیق رکوردهای ثبت شده از شبکه صداوسیمای مرکز بوشهر است که از طریق گزارش‌گیری از داده‌های ورودی شبکه استخراج شده است. علاوه بر این، نتایج شبیه‌سازی نشان می‌دهد که روش پیشنهادی با دقت میانگین ۹۷٫۶۴٪ روی مجموعه داده NSL-KDD عملکرد مناسبی را در تشخیص نفوذ و ارتقاء امینت فراهم می‌کند.

کلیدواژه‌ها: *تشخیص نفوذ، ارتقای امنیت، شبکه صداوسیمای مرکز بوشهر، الگوریتم PCA، طبقه‌بندی ترکیبی*

مقدمه

تکنولوژی اینترنت به‌طور چشمگیری ارتباط با دیگران را تغییر داده است. شبکه‌های کامپیوتری یکی از سرویس‌های که امروزه اینترنت برای ارتباطات و انتقال اطلاعات بین کاربران فراهم کرده است. امروزه تهدیداتی که از اینترنت می‌آیند بیش‌ازپیش پیچیده شده و توانایی این را دارند که از راه‌حل‌های امنیتی معمول از قبیل دیوار آتش و آنتی‌ویروس نیز رد شوند. امروزه، نفوذ به یک نگرانی فزاینده تبدیل شده است[۱]. با ظهور فناوری‌های جدید هرروز و گسترش رایانه‌ها، امنیت به موضوعی بسیار مهم تبدیل شده است. ازاین‌رو، امنیت هر شبکه یک موضوع مهم با اولویت است که باید از آن مراقبت کرد.

یک‌راه حل ممکن برای بهبود امنیت، اضافه کردن یک سیستم تشخیص نفوذ است که به‌عنوان یک‌لایه اضافی در راه‌حل‌های امنیتی محسوب می‌شود. این سیستم‌ها، مؤلفه اصلی یک شبکه امن هستند. سیستم‌های تشخیص نفوذ می‌توانند به شناسایی تهدیدات در شبکه‌ها کمک کنند و یک ابزار لازم و مهم برای زیرساخت شبکه‌ها می‌باشند [۲]. از تشخیص نفوذ برای شناسایی مزاحم‌ها و کاربران خود سیستم که از امتیازات قانونی خود سوء استفاده می‌کنند استفاده می‌شود. سیستم‌های تشخیص نفوذ سنتی نمیتوانند خود را با حملات جدید تطبیق دهند. ازاین‌رو، امروزه سیستم‌های تشخیص نفوذ مبتنی بر داده کاوی مطرح شده است. روش‌های داده کاوی با مشخص نمودن یک برچسب دودویی (بسته نرمال، بسته غیرنرمال) و همچنین مشخص نمودن ویژگی‌ها و خصیصه با الگوریتم‌های انتخاب ویژگی می‌توانند داده غیر نرمال به خوبی تشخیص دهند[۳]. بنابراین، دقت و درستی سیستم‌های تشخیص نفوذ افزایش یافته و در نتیجه امنیت شبکه بالا می‌رود. طبقه‌بندی یکی از مهمترین و گسترده‌ترین تکنیک‌های با ناظر در داده کاوی است. این یک فرآیند دو مرحله‌ای است و داده‌های ورودی را به دو دسته تقسیم می‌کند؛ در مرحله اول (مرحله یادگیری)، الگوریتم طبقه‌بندی داده‌ها را با استفاده از مدل‌سازی تجزیه و تحلیل می‌کند و به بررسی روابط و الگوها می‌پردازد. این در حالیست که در مرحله دوم از مدل ایجاد شده برای کار تشخیص و پیش‌بینی داده‌های جدید استفاده می‌شود.

در حالیکه شبکه‌های صدا و سیما یک‌راه‌حل امیدوارکننده برای حل کمبود طیف رسانه‌ای ارائه می‌دهند، آنها هنوز هم در معرض تهدیدات امنیتی قرار دارند. تاکنون فقط تعداد معدودی محقق استفاده از سیستم‌های تشخیص نفوذ را برای مقابله با این تهدیدات علیه شبکه‌های صدا و سیما در نظر گرفته‌اند[۴]. سیستم تشخیص نفوذ یکی از تکنیک‌های گسترده است که در یک توپولوژی شبکه به کار میرود تا از یکپارچگی و در دسترس بودن دارایی‌های حساس در سیستم‌های امن محافظت کند. اگرچه بسیاری از رویکردهای یادگیری با ناظر و بدون ناظر از حوزه یادگیری ماشین برای افزایش اثربخشی سیستم تشخیص نفوذ استفاده شده است، اما هنوز هم برای دستیابی به عملکرد خوب، الگوریتم‌های تشخیص نفوذ نفوذ باید بهبود یابند[۵]. از چالش‌های بهبود سیستم‌های تشخیص نفوذ میتوان به موارد زیر اشاره کرد: ۱) تعداد زیادی از داده‌های زائد و بی‌ربط در مجموعه داده‌های مرتبط با روند طبقه‌بندی یک سیستم تشخیص نفوذ تداخل دارند. ۲) ممکن است یک مدل طبقه‌بندی انفرادی در تشخیص یک نوع از حملات عملکرد خوبی نداشته باشد. ۳) بسیاری از مدل‌ها برای مجموعه داده‌های قدیمی طراحی شده‌اند و این منجر به کاهش سازگاری آنها برای حملات جدید می‌شود[۶].

یکی از بزرگترین مشکلات و چالش‌ها در جهت بهبود امنیت در شبکه‌های کامپیوتری سازمان‌ها، مربوط به شناسایی و تشخیص مؤثر و مناسب حملات سایبری و نفوذهای غیر مجاز می‌باشد. با توجه به تلاش‌های زیادی که در پژوهش‌های گذشته به موضوع تشخیص نفوذ پرداخته‌اند، هنوز هم این مبحث به دلیل ضرورت امنیتی مورد توجه می‌باشد[۷]. در این تحقیق یک رویکرد ترکیبی برای شناسایی نفوذ و حملات اینترنتی به شبکه‌های کامپیوتر سازمان صدا و سیما ارائه شده است. روش پیشنهادی شامل یک الگوریتم PCA برای تشخیص ویژگی‌های مؤثر و یک روش طبقه‌بندی ترکیبی برای مدل‌سازی داده‌های نفوذ است. در این مقاله از مجموعه داده NSL-KDD و همچنین پایگاه داده رکورد شده از شبکه صداوسیمای مرکزی استان بوشهر برای شبیه‌سازی روش پیشنهادی استفاده شده است.

ادامه این مقاله به شرح زیر است: برخی از جدیدترین کارهای انجام شده در بخش ۲ بررسی شده است. در بخش ۳ روش پیشنهادی مبتنی بر الگوریتم PCA و رویکرد طبقه‌بندی ترکیبی به منظور تشخیص ارائه می‌شود. نتایج حاصل از ارزیابی روش پیشنهادی برای تشخیص نفوذ در رکوردهای شبکه صدا و

سیما استان بوشهر در بخش ۴ آورده شده و در نهایت نتیجه-گیری و پیشنهادها در بخش ۵ بیان شده است.

پیشینه تحقیق

در این بخش درباره کارهای صورت گرفته در حوزه تشخیص نفوذ بحث می‌شود. در[۸] یک سیستم تشخیص نفوذ برای مبارزه با حملات علیه شبکه‌های رادیویی پیشنهاد شد. در این مقاله شبکه‌های رادیویی بر اساس شبکه منطقه‌ای بی‌سیم و برخی از تهدیدات امنیتی علیه آن توصیف شده است. برای اینکه کاربران ثانویه مورد حمله در شبکه‌های رادیویی به سرعت تشخیص داده شوند، یک شناسه ساده اما مؤثر ارائه شده است. پیشنهاد این تحقیق از میزان تجمع غیر پارامتری به عنوان الگوریتم تشخیص نقطه تغییر برای کشف رفتار غیرعادی ناشی از حملات است. سیستم تشخیص نفوذ پیشنهاد شده یک روش تشخیص ناهنجاری را اتخاذ کرده و پارامترهای سیستم شبکه-های رادیویی را از طریق یک مرحله یادگیری پروفایل می‌کند. در[۹] رویکرد یادگیری نیمه ناظار مبتنی بر ترکیب برای یک سیستم تشخیص نفوذ توزیع شده پیشنهاد شده است. روشی که برای انتخاب ویژگی‌ها استفاده شده براساس شاخص جینی و اندازه‌گیری اهمیت متغیر است. طبقه‌بندی‌های مورد استفاده نیز جنگل تصادفی، ماشین بردار پشتیبان، شبکه عصبی مصنوعی، درخت تصمیم‌گیری و K نزدیکترین همسایه می‌باشند. برای ترکیب طبقه‌بندها از یک اسمبلینگ برای دادن کلاس نهایی از چند طبقه‌بند استفاده شده است.

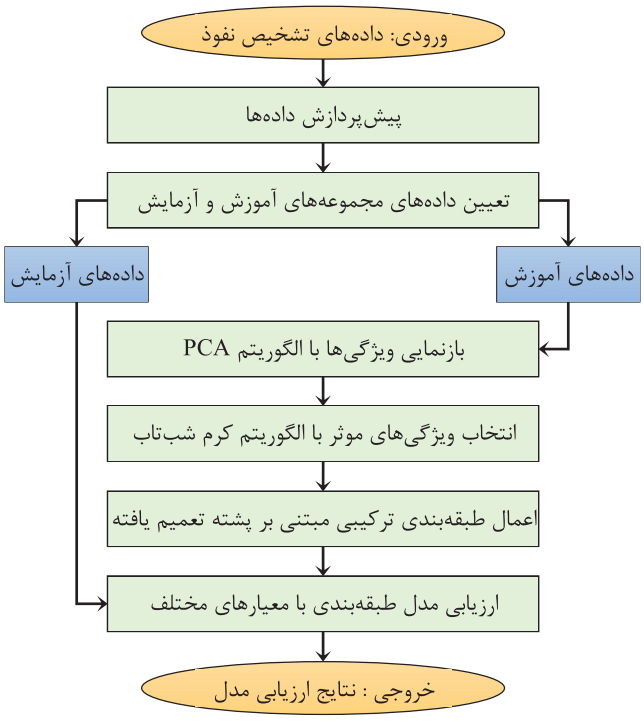
در[۱۰] یک روش جدید برای تشخیص نفوذ با استفاده از ترکیب ماشین بردار پشتیبان با تقویت ویژگی ارائه شده است. در این مقاله، تبدیل نسبت‌های چگالی حاشیه‌ای لگاریتم روی ویژگی-های اصلی با هدف به دست آوردن داده‌های آموزش تبدیل شده با کیفیت بهتر انجام می‌شود. سپس از مجموعه ماشین بردار پشتیبان برای ساخت مدل تشخیص نفوذ استفاده می‌شود. در [۱۱]کاهش ابعاد با IG-PCA و طبقه‌بندی ترکیبی برای تشخیص نفوذ شبکه معرفی شده است. در این مقاله یک تکنیک کاهش ابعاد ترکیبی جدید برای شناسایی نفوذ، ترکیب رویکردهای افزایش اطلاعات و تجزیه و تحلیل مؤلفه اصلی با استفاده از یک طبقه‌بندی گروهی مبتنی بر ماشین بردار پشتیبان، الگوریتم‌های یادگیری مبتنی بر نمونه و پرسپترون چند لایه پیکربندی شده است. در [۱۲] مدل تشخیص نفوذ یکپارچه با استفاده از انتخاب ویژگی chi-square و طبقه‌بندی ترکیبی معرفی شد. هدف از این مقاله شناسایی ویژگی‌های مهم مورد نیاز در ساخت مدل تشخیص نفوذ و در نتیجه دستیابی به حداکثر دقت است. این مدل از یک رویکرد طبقه‌بندی گروهی با حداقل پیچیدگی برای غلبه بر مسائل موجود در مدل‌های تشخیص نفوذ استفاده می‌کند. در این مقاله، از انتخاب ویژگی-های Chi-square و مدل‌های طبقه‌بندی ماشین بردار پشتیبان، نائو بیز و غیره برای ایجاد یک مدل تشخیص نفوذ ترکیبی استفاده شده است. از آنجا که یک عدم تعادل کلاس عظیم در ترافیک شبکه وجود دارد، پیش بینی برچسب کلاس با رأی اکثریت یک‌راه حل بهینه در اولویت اعتماد به یک طبقه‌بندی واحد است.

الگوریتم [۱۳] ارائه شده است. این الگوریتم یک سیستم تشخیص نفوذ مؤثر بر اساس انتخاب ویژگی و طبقه‌بندی ترکیبی است. این روش ابتدا زیرمجموعه ویژگی‌های بهینه را بر اساس همبستگی انتخاب می‌کند. سپس، از الگوریتم‌های C4.5، RF و PA برای ایجاد یه مدل طبقه‌بندی ترکیبی استفاده می‌کند. در نهایت، از روش رای گیری اکثریت را برای ترکیب توزیع احتمال طبقه‌بندی‌ها بهره می‌گیرد. الگوریتم TSE-IDS [۱۴] ارائه شده است. این روش یک طبقه‌بندی ترکیبی دو مرحله‌ای مبتنی بر تشخیص ویژگی‌های ناهنجار است. تکنیک انتخاب ویژگی ترکیبی است و شامل PSO، ACO و GA است. در اینجا کارایی ویژگی‌ها بر اساس مدل طبقه‌بندی درخت هرس خطا انجام شده است و در نهایت، یک مدل طبقه‌بندی ترکیبی در دو سطح برای مدل‌سازی داده‌های تشخیص نفوذ اعمال می‌گردد. الگوریتم ELNIDS [۱۵] ارائه شده است. این روش یک سیستم یادگیری ترکیبی در شبکه مبتنی بر اینترنت چیزها است. در اینجا از چهار طبقه‌بندی مختلف ترکیبی شامل درخت تصمیم تقویت شده، درخت تصمیم شکسته شده، درخت تصمیم تبعیض آمیز و درخت تصمیم RUSBoosted برای تشخیص نفوذ استفاده شده است.

روش پیشنهادی

روش پیشنهادی شامل چندین گام است: در ابتدا داده‌های مربوط به نفوذ از مجموعه داده ورودی پیش پردازش می‌شوند تا

از کیفیت لازم برای آموزش مدل برخوردار باشند. سپس داده‌های مربوط به مجموعه‌های آموزش و آزمایش مشخص می‌شوند. در این مقاله، از یک الگوریتم کرم شب تاب که با PCA توسعه یافته برای انتخاب زیرمجموعه ویژگی‌هایی مؤثر در شناسایی نفوذ استفاده می‌شود. روش انتخاب ویژگی پیشنهادی مبتنی بر اندازه گیری توزیع ویژگی‌ها و توزیع داده‌ها است. به منظور بهبود الگوریتم کرم شب تاب در انتخاب ویژگی‌ها، از خصوصیات الگوریتم PCA استفاده شده که میتواند روند انتخاب ویژگی را بهبود بخشد، زیرا انتخاب بر اساس PCA مانع از انتخاب ویژگی‌های مشابه که تأثیر یکسانی در شناسایی نفوذ دارد، می‌شود. در مرحله بعد، از یک مدل طبقه‌بندی ترکیبی برای مدل‌سازی داده‌های نفوذ و محاسبه تناسب هر راه‌حل استفاده می‌شود. بنابراین، این مرحله از روش پیشنهادی به ارائه یک روش طبقه‌بندی ترکیبی برای محاسبه تناسب هر راه‌حل از الگوریتم کرم شب‌تاب اختصاص دارد. روش طبقه‌بندی ترکیبی از k مدل طبقه‌بندی کلاسیک نظیر SVM، ID3 و KNN استفاده می‌کند و از تکنیک پشته تعمیم یافته برای ترکیب خروجی‌ها بهره می‌گیرد. شکل ۱ فلوچارت روش پیشنهادی را نشان می‌دهد.



شکل ۱: فلوچارت روش پیشنهادی

پیش پردازش داده‌ها

در این مرحله از تکنیک‌های پاک‌سازی و نرمال‌سازی استفاده می‌شود. در پاک‌سازی نمونه‌هایی که دارای مقادیر گم شده (مقادیر از دست رفته با نماد «؟») هستند از مجموعه داده حذف شده و با تخصیص مقادیر ثابت منحصر به فرد، ویژگی‌های کیفی را به کمی تبدیل می‌کنیم. در نرمال‌سازی نگاشت محدوده ارزش ویژگی‌ها به یک محدوده مشخص و یکسان انجام می‌شود. در اینجا، نرمال‌سازی بر اساس رابطه (۱) انجام شده است.

$$x_{i,j}^{norm} = \frac{x_{i,j}}{\max\limits_{k=1..N}(k,j)} \qquad (۱)$$

در این رابطه، $x_{i,j}^{norm}$ و $x_{i,j}$ به ترتیب مقدار ویژگی نرمال شده و واقعی برای نمونه i و ویژگی j است و N تعداد کل نمونه‌ها را نشان می‌دهد.

تعیین داده‌های آموزش و آزمایش

به‌طور کلی داده‌های بخش آموزش برای ساخت مدل و از داده‌های بخش آزمایش برای تست و ارزیابی مدل استفاده می‌شود. در این تحقیق از اعتبارسنجی 10-fold برای تعیین داده‌های آموزش و آزمایش استفاده شده است. در این روش همه نمونه‌ها از مجموعه داده به ۱۰ بخش با اندازه برابر افراز شده و از ۹ بخش برای کار آموزش و از بخش باقیمانده برای کار آزمایش مدل استفاده می‌شود. این فرایند ۱۰ بار تکرار شده بطوریکه همه بخش‌ها یکبار برای آزمایش استفاده شوند. در نهایت میانگین ارزیابی برای ۱۰ اجرای مستقل به عنوان خروجی گزارش می‌شود.

بازنمایی ویژگی‌ها با الگوریتم PCA

در اینجا از الگوریتم تحلیل مولفه‌های اصلی (PCA) برای بازنمایی ویژگی‌ها و در نهایت کاهش ابعاد داده‌ها استفاده می‌کنیم. بر مبنای PCA، چرخاندن دستگاه مختصات در جهت بیشترین واریانس داده‌ها میتواند در کاهش تعداد ویژگی‌ها مؤثر باشد. بیشترین واریانس داده‌ها در راستای بردار ویژه متناظر با بزرگترین مقدار ویژه است. ازاین‌رو هرچه مقدار ویژه کمتر باشد واریانس داده‌ها در راستای بردار ویژه متناظر با آن نیز کمتر است. با اعمال PCA روی مجموعه داده ورودی، تعداد K مؤلفه

اصلی از مجموعه داده با ابعاد $N \times D$ استخراج می‌شود. بنابراین مجموعه داده به ابعاد $N \times K$ نگاشت داده می‌شود. در اینجا، N تعداد نمونه‌ها، D تعداد ویژگی‌های اصلی و K تعداد مولفه‌های اصلی است. با ترکیب ویژگی‌های اصلی و استخراج شده، ساختار جدید مجموعه داده بصورت $N \times M$ جائیکه $M = D + K$ می‌باشد.

انتخاب ویژگی با الگوریتم کرم شب تاب

اکنون با استفاده از الگوریتم کرم شب تاب [۱۶]، زیرمجموعه ویژگی‌های مؤثر در تشخیص نفوذ شناسایی می‌شوند. در اینجا ساختار نمایش کرم‌های شب تاب (راه‌حل‌ها) به صورت یک بردار باینری M بیتی است که برای هر اندیس ۰ عدم انتخاب و ۱ بیانگر انتخاب ویژگی مربوطه است. در اینجا M تعداد کل ویژگی‌ها بر مبنای ترکیب مجموعه داده اصلی و الگوریتم PCA است. جمعیت اولیه نیز به صورت تصادفی با توجه به ساختار نمایش کرم‌ها تولید می‌شود. اگر اندازه جمعیت N_p باشد، بنابراین جمعیت اولیه ماتریسی در ابعاد $N_p \times M$ است. در این تحقیق لوسی‌فرین هر کرم شب‌تاب، دقت طبقه‌بندی ویژگی‌های انتخاب شده از هر راه‌حل را با توجه به مجموعه داده آموزشی نشان می‌دهد. $\ell(i)$ لوسی‌فرین کرم شب‌تاب i-ام است و با استفاده از یک رویکرد طبقه‌بندی ترکیبی محاسبه می‌شود.

احتمال حرکت برای کرم شب تاب i-ام نشان می‌دهد که شب تاب i-ام با چه احتمالی به سمت همسایگان خود حرکت می‌کند. برای تشخیص همسایگان کرم شب تاب i-ام از پارامتر شعاع حسگری (rs) استفاده می‌شود. در اینجا، rs با توجه به مقدار لوسی‌فرین تعیین می‌گردد. در این مقاله همسایگان کرم شب تاب i-ام، کرم‌های شب‌تابی از همه جمعیت با مقدار لوسی‌فرین بیشتر هستند. در الگوریتم کرم شب‌تاب شعاع تصمیم (rd) نشان‌دهنده برد حسگری هر کرم شب تاب نسبت به کل فضای جستجو می‌باشد. در اینجا، همسایگان کرم شب‌تاب i-ام (N_i)، تمامی کرم‌های شب‌تابی هستند که مقدار لوسی‌فرین آنها حداکثر به مقدار آستانه th از مقدار لوسی‌فرین کرم شب‌تاب i-ام بیشتر است. اکنون احتمال حرکت کرم شب‌تاب i-ام به سمت j-مین کرم شب‌تاب از مجموعه N_i به صورت رابطه (۲) محاسبه می‌شود.

$$p_{ij} = \frac{\ell(j) - \ell(i)}{\sum_{k \in N_i} \ell(k) - \ell(i)} \qquad (۲)$$

در نهایت برای کرم شب تاب i-ام، یک کرم شب تاب از مجموعه N_i با استفاده از ساختار چرخ رولت انتخاب می‌شود. برای شبیه‌سازی حرکت کرم شب تاب i-ام به سمت کرم شب تاب j-ام از رابطه (۳) استفاده می‌شود.

$$X_i(k) = \begin{cases} X_i(k) & \alpha > R \\ X_j(k) & Otherwise \end{cases}, k = 1,2,\dots,M \qquad (۳)$$

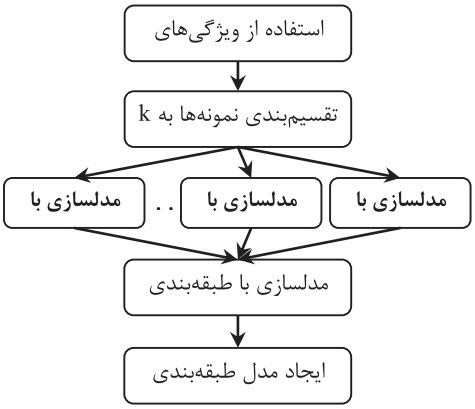
در این رابطه، R یک عدد تصادفی بین $[0,1]$، α پارامتر کنترلی ثابت به منظور همگرایی الگوریتم، $X_i(k)$ موقعیت k-ام از کرم شب‌تاب i-ام و $X_j(k)$ موقعیت k-ام از کرم شبت‌تاب j-ام می‌باشد. برای بروزرسانی لوسی‌فرین کرم شب‌تاب i-ام از مقدار قبلی آن و همچنین مقدار تابع برازندگی مطابق رابطه (۴) استفاده می‌شود.

$$\ell_i(t) = (1-\rho)\,\ell_i(t-1) + \gamma\, J(X_i(t)) \qquad (۴)$$

در این رابطه، $\ell_i(t)$ و $\ell_i(t-1)$ به ترتیب مقدار جدید و قبلی لوسی‌فرین برای کرم شب‌تاب i-ام می‌باشد. $J(X_i(t))$ مقدار برازندگی کرم شب‌تاب i-ام را در تکرار t (تکرار جاری) نشان می‌دهد. به منظور مدل کردن افت تدریجی با زمان مقداری از لوسی‌فرین فعلی با ضریبی کمتر از ۱ توسط پارامتر کنترلی ρ کم می‌شود. پارامتر γ یک ثابت عددی برای مدل کردن تاثیر برازندگی بر لوسی‌فرین است. حداکثر تکرار مراحل الگوریتم کرم شب تاب با توجه به پارامتر Max_{Iter} تعیین می‌شود.

رویکرد طبقه‌بندی ترکیبی

در اینجا، از یک رویکرد طبقه‌بندی ترکیبی برای مدل‌سازی داده‌های آموزشی و محاسبه برازندگی راه‌حل‌ها در الگوریتم کرم شب تاب استفاده می‌شود. نمونه‌های آموزشی به k دسته با اندازه برابر تقسیم شده و سپس برای مدل‌سازی و آموزش هر دسته از یک روش طبقه‌بندی انفرادی استفاده می‌شود. خروجی روش‌های طبقه‌بندی انفرادی بر اساس تکنیک پشته تعمیم یافته ادغام می‌گردند. این فرایند آموزش به عنوان روش bagging [۱۷] شناخته می‌شود. شکل ۲ معماری رویکرد طبقه‌بندی ترکیبی پیشنهادی را نشان می‌دهد.



شکل ۲: معماری رویکرد طبقه‌بندی ترکیبی

در اینجا طبقه‌بندی ترکیب یا فرا-طبقه‌بندی کار یادگیری نتایج پیش‌بینی‌ها در روش‌های طبقه‌بندی انفرادی بر اساس برچسب خروجی واقعی را انجام می‌دهد. در واقع ورودی فرا-طبقه‌بندی، برچسب پیش‌بینی و برچسب واقعی از k روش طبقه‌بندی انفرادی است. این فرایند ادغام خروجی‌ها مطابق تکنیک پشته تعمیم یافته است که بجای رای اکثریت از یادگیری رفتار روش-های طبقه‌بندی انفرادی استفاده می‌کند.

نتایج و مقایسه‌ها

در این بخش آزمایش‌های گسترده‌ای برای بررسی عملکرد الگوریتم پیشنهادی انجام شده است. شبیه‌سازی با نرم‌افزار متلب ورژن 2019a انجام شده و همه آزمایش‌ها بوسیله یک PC با پردازنده اینتل core i7 و فرکانس 3.2GHz، حافظه 16GB اعمال می‌شود. در آزمایش‌های انجام شده مقادیر تنظیم شده برای پارامترهای روش پیشنهادی به صورت زیر است.

$N_p = 75,\ th = 0.15,\ \alpha = 0.4,\ \rho = 0.25,\ \gamma = 0.2,$
$Max_{Iter} = 200,\ k = 3$

در اینجا از SVM، ID3 و KNN به عنوان مدل‌های طبقه‌بندی انفرادی و از لجستیک رگرسیون به عنوان مدل طبقه‌بندی ترکیب استفاده شده است. علاوه بر این، از مجموعه داده NSL-KDD و همچنین پایگاه داده رکوردهای شده از شبکه صداوسیمای مرکزی استان بوشهر برای شبیه‌سازی و مقایسه استفاده شده است. مجموعه داده NSL-KDD این مجموعه داده شامل ۴۱ ویژگی، ۲۵۱۹۳ نمونه در دو کلاس نرمال و نفوذ است. کلاس نفوذ نیز دارای چهار زیر کلاس Dos، R2L، U2R و Probe می‌باشد. مجموعه داده صدا و سیما مرکز بوشهر به صورت دستی از ثبت رکوردهای ورودی به شبکه صدا و سیما استان بوشهر استخراج شده است که شامل ۳۰۰ نمونه و ۱۱ ویژگی در ۲ کلاس نرمال و نفوذ است. ویژگی‌های این مجموعه داده شامل ۱- تاریخ (Date)، ۲- ساعت (Time)، ۳- آدرس IP مبدا (SRC IP)، ۴- آدرس پورت مبدا (SRC port)، ۵- آدرس IP مقصد (DST IP)، ۶- آدرس پورت مقصد (DST port)، ۷- تعداد درخواست‌های متوالی (Req)، ۸- نوع سرویس درخواست شده (Service)، ۹- نوع پروتکل مورد استفاده (Protocol)، ۱۰- دامنه (Domain) و ۱۱- تعداد بایت‌های مورد دسترسی (Byte-Length) می‌باشد.

در آزمایش اول نتایج روش پیشنهادی و مدل‌های طبقه‌بندی ID3، SVM و CART و همچنین الگوریتم‌های CFS-BA، TSE-IDS و ELNIDS از لحاظ معیارهای مختلف روی مجموعه داده NSL-KDD مقایسه شده است. این معیارها شامل دقت، درستی، فراخوان، اندازه‌گیری F و زمان اجرا مطابق جدول ۱ است. نتایج برتری روش پیشنهادی را نسبت به الگوریتم‌های ID3، SVM و CART به ترتیب با ۳۸,۹۴٪، ۳۲,۳۰٪ و ۳۷,۴۲٪ در معیار دقت نشان می‌دهد. این برتری نسبت به الگوریتم‌های CFS-BA، TSE-IDS به ترتیب ۰,۲۴٪ و ۸,۶٪ می‌باشد، این درحالیست که و ELNIDS نسبت به روش پیشنهادی ۲,۱٪ برتری دارد. دلیل برتری سایر مدل‌های طبقه-بندی از روش پیشنهادی در معیار زمان اجرا، این است که روش پیشنهادی حاوی یک الگوریتم بهینه‌سازی برای انتخاب زیرمجموعه ویژگی‌های بهینه است، ولی مدل‌های طبقه‌بندی از همه ویژگی‌ها برای کار مدل‌سازی استفاده می‌کنند.

جدول ۱. مقایسه روش پیشنهادی با مدل‌های طبقه‌بندی مختلف روی مجموعه داده NSL-KDD

روش‌ها	دقت (٪)	درستی (٪)	فراخوان (٪)	اندازه‌گیری F (٪)	زمان اجرا (s)
ID3	۷۰,۲۷	۵۸,۹۹	۸۴,۰۳	۶۹,۳۲	۷۳۳
SVM	۷۳,۸۰	۶۹,۸۷	۷۴,۴۳	۷۲,۰۸	۱۰۱۸
CART	۷۱,۰۵	۶۶,۲۳	۷۴,۱۹	۶۹,۹۸	۸۱۶
CFS-BA	۸۹,۹۱	۸۱,۵۰	۹۲,۱۲	۸۹,۳۴	۷۸۴۳
TSE-IDS	۹۷,۴۱	۹۷,۱۷	۹۵,۳۱	۹۷,۰۵	۱۰۲۲۷
ELNIDS	۹۹,۷۱	۹۷,۷۵	۹۸,۱۲	۹۹,۶۱	۶۲۰۴

روش پیشنهادی	۹۷,۶۴	۹۸,۲۳	۹۵,۸۶	۹۷,۰۵	۵۸۱۷

در آزمایش دوم مقایسه‌ای بین روش پیشنهادی و سایر روش-های مورد مقایسه در دو حالت بدون انتخاب ویژگی و انتخاب ویژگی انجام شده است. این مقایسه برای هر دو مجموعه داده NSL-KDD و صدا و سیما در جدول ۲ گزارش شده است.

جدول ۲: مقایسه نتایج روش پیشنهادی با سایر روش‌ها روی مجموعه داده NSL-KDD و صدا و سیما

روش‌ها	مجموعه داده	دقت بدون انتخاب ویژگی (٪)	دقت با انتخاب ویژگی (٪)
ID3	NSLKDD	۷۰/۲۷	-
	صدا و سیما	۸۵/۱۳	-
SVM	NSLKDD	۷۳/۸۰	-
	صدا و سیما	۸۹/۵۱	-
CART	NSLKDD	۷۱/۰۵	-
	صدا و سیما	۸۷,۱۶	-
CFS-BA	NSLKDD	۹۶/۹۱	۸۹/۹۱
	صدا و سیما	-	-
TSE-IDS	NSLKDD	۹۷/۸۶	۹۷/۴۱
	صدا و سیما	-	-
ELNIDS	NSLKDD	۹۹/۶۸	۹۹/۷۱
	صدا و سیما	-	-
روش پیشنهادی	NSLKDD	۹۷/۴۸	۹۷/۶۴
	صدا و سیما	۹۸/۲۵	۹۸/۳۴

نتیجه‌گیری و پیشنهادها

سیستم‌های تشخیص نفوذ به‌طور گسترده برای محافظت از امنیت شبکه استفاده می‌شوند. تکنیک‌های مختلف یادگیری ماشین برای بهبود عملکرد این سیستم‌ها بسیار رایج هستند و از میان آنها، یادگیری ترکیبی توجه فزاینده‌ای پیدا کرده و به عنوان روشی مؤثر در نظر گرفته می‌شود. در این مقاله یک رویکرد ترکیبی شامل الگوریتم PCA برای انتخاب ویژگی‌ها و طبقه‌بندی ترکیبی برای کار مدل‌سازی داده‌ها به منظور تشخیص نفوذ روی شبکه صداوسیمای مرکزی استان بوشهر ارائه شد. رسیدگی به ویژگی‌های نامربوط در مجموعه داده‌های بعد بالا باعث یک چالش طولانی‌مدت برای تشخیص ناهنجاری شبکه شده است. از بین بردن چنین ویژگی‌هایی با اطلاعات طیفی نه تنها سرعت فرآیند طبقه‌بندی را بالا می‌برد بلکه به مدل‌های طبقه‌بندی کمک می‌کند که در طول زمان تشخیص حمله، تصمیم‌های دقیق بگیرند، به خصوص زمانی که با داده‌های با مقیاس بزرگ و ناهمگن مواجه هستند. در اینجا الگوریتم کرم شب‌تاب با استفاده از الگوریتم PCA توانسته به خوبی داده‌های زائد را تشخیص و حذف کند. این عما در نهایت منجر به بهبود فرایند مدل‌سازی توسط رویکرد طبقه‌بندی ترکیبی شده است. برای کارهای آینده پیشنهاد می‌شود از سایر الگوریتم‌های تکاملی نظیر ژنتیک و پرندگان به جای الگوریتم کرم شب‌تاب برای بهینه‌سازی ویژگی‌های استفاده شود. علاوه بر این، می‌توان با ادغام چندین الگوریتم تکاملی فرایند انتخاب ویژگی‌ها را نیز به صورت ترکیبی و مطلوب‌تر انجام داد.

منابع

[1] Salo, F., Nassif, A. B., & Essex, A. (2019). Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. *Computer Networks*, *148*, 164-175.

[2] Jabbar, M. A., & Aluvalu, R. (2017). RFAODE: A novel ensemble intrusion detection system. *Procedia computer science*, *115*, 226-234.

[3] Vinutha, H. P., & Poornima, B. (2018). An ensemble classifier approach on different feature selection methods for intrusion detection. In *Information systems design and intelligent applications* (pp. 442-451). Springer, Singapore.

[4] Jabbar, M. A., Srinivas, K., & Reddy, S. S. S. (2016, December). A Novel Intelligent Ensemble Classifier for Network Intrusion Detection System. In *International Conference on Soft Computing and Pattern Recognition* (pp. 490-497). Springer, Cham.

[5] Aburomman, A. A., & Reaz, M. B. I. (2016). A novel SVM-kNN-PSO ensemble method for intrusion detection system. *Applied Soft Computing*, *38*, 360-372.

[6] Ghalehgolabi, M., & Rezaeipanah, A. (2017). Intrusion Detection System Using Genetic

Algorithm and Data Mining Techniques Based on the Reduction. *International Journal of Computer Applications Technology and Research*, *6*(11), 461-466.

[7] Subasi, A., Qaisar, S. M., Al-Nory, M., & Rambo, K. A. (2019). Intrusion Detection in Smart Grid Using Bagging Ensemble Classifiers. In *2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)* (pp. 1-5). IEEE.

[8] Fadlullah, Z. M., Nishiyama, H., Kato, N., & Fouda, M. M. (2013). Intrusion detection system (IDS) for combating attacks against cognitive radio networks. *IEEE network*, *27*(3), 51-56.

[9] Zhou, Y., Cheng, G., Jiang, S., & Dai, M. (2020). Building an efficient intrusion detection system based on feature selection and ensemble classifier. *Computer Networks*, 107247.

[10] Tama, B. A., Comuzzi, M., & Rhee, K. H. (2019). TSE-IDS: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system. *IEEE Access*, *7*, 94497-94507.

[11] Salo, F., Nassif, A. B., & Essex, A. (2019). Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. *Computer Networks*, *148*, 164-175.

[12] Thaseen, I. S., Kumar, C. A., & Ahmad, A. (2019). Integrated intrusion detection model using chi-square feature selection and ensemble of classifiers. *Arabian Journal for Science and Engineering*, *44*(4), 3357-3368.

[13] Verma, A., & Ranga, V. (2019, April). ELNIDS: Ensemble learning based network intrusion detection system for RPL based Internet of Things. In *2019 4th International conference on Internet of Things: Smart innovation and usages (IoT-SIU)* (pp. 1-6). IEEE.

[14] Khonde, S. R., & Ulagamuthalvi, V. (2019). Ensemble-based semi-supervised learning approach for a distributed intrusion detection system. *Journal of Cyber Security Technology*, *3*(3), 163-188.

[15] Gu, J., Wang, L., Wang, H., & Wang, S. (2019). A novel approach to intrusion detection using SVM ensemble with feature augmentation. *Computers & Security*, *86*, 53-62.

[16] Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, *36*(1-2), 105-139.

[17] Yang, X. S. (2010). Firefly algorithm, stochastic test functions and design optimisation. *International journal of bio-inspired computation*, *2*(2), 78-84.