بررسی روش‌های کاهش ابعاد در سیستم‌های تشخیص نفوذ

رویا گلستانی شیشوان [*۱]، امیرفرید امینیان مدرس[۲].

۱- دانشجوی کارشناسی ارشد مهندسی کامپیوتر نرم‌افزار، دانشگاه صنعتی سجاد، مشهد

۲- عضو هیئت علمی، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی سجاد، مشهد

چکیده

برای ایجاد امنیت داده‌ها در یک سیستم کامپیوتری، علاوه بر دیوارهای آتش و دیگر تجهیزات جلوگیری از نفوذ، سیستم‌های دیگری به نام تشخیص نفوذ (IDS) مورد نیاز می‌باشند تا بتوانند در صورتی که نفوذگر از دیوار آتش و دیگر تجهیزات امنیتی عبور کرد، آن را تشخیص دهد. تشخیص نفوذ، با مشکلاتی از قبیل حجم عظیم ترافیک شبکه، توزیع داده های بسیار نامتعادل، سختی شناسایی مرزهای تصمیم‌گیری بین رفتار طبیعی و غیرطبیعی، مواجه است. در این مقاله، هفت روش تبدیلی کاهش ابعاد (شامل دو نوع خطی و غیرخطی) برای سیستم تشخیص نفوذ مبتنی بر ماشین بردار پشتیبان، مقایسه و ارزیابی می‌شوند. این روش‌ها شامل تحلیل مولفه‌های اصلی ، تحلیل متمایز خطی (LDA)، تحلیل عاملی (FA)، تحلیل مولفه اساسی مبتنی بر هسته (Kernel PCA)، کدکننده‌های خودکار(Autoencoder)، تحلیل مولفه های اصلی احتمالاتی (ProbPca) و مقیاس‌گذاری چندبعدی (MDS) می‌باشند. در این مقاله از مجموعه داده‌ی NSL-KDD جهت ارزیابی سیستم تشخیص نفوذ و روش‌های کاهش ابعاد استفاده می‌شود. همچنین مدت زمان آموزش، آزمایش، میزان صحت، دقت و درصد خطا توسط سیستم تشخیص نفوذ باهم مقایسه می‌شود. نتایج مقایسه‌ها نشان می‌دهد که روش ProbPca زمان آموزش و آزمایش مناسب‌تری از روش‌های تبدیلی دیگر دارد.

کلمات کلیدی: تشخیص نفوذ، کاهش ابعاد، امنیت داده‌ها.

۱. مقدمه

رشد تکنولوژی اینترنت، ارتباطات داده‌ها را گسترش داده است. از طرفی تهدیدات شبکه و مسائل امنیتی، مشکل جامعیت داده‌ها و از دست رفتن اطلاعات را افزایش می‌دهد. با توجه به این موضوع بحث امنیت اهمیت بیشتری پیدا کرده است. تعریف امنیت داده‌ها شامل تعریف سه بخش: یکپارچگی، محرمانگی و دسترس‌پذیری داده‌ها است. یکپارچگی داده‌ها به این معناست که تنها افراد و سیستم‌های مجاز می‌توانند در داده تغییر ایجاد کنند. محرمانگی داده به این معناست که تنها افراد و سیستم‌های مجاز می‌توانند اطلاعات حساس یا طبقه‌بندی شده را ببینند و سایر موارد امکان هیچگونه دسترسی به

* Email: golestani.roya@gmail.com

این داده‌ها را نداشته باشند. دسترس‌پذیری شامل داده‌ها و سیستم‌ها می‌شود. اگر شبکه یا داده آن، برای کاربران مجاز در دسترس نباشد (به دلایلی مانند حمله DoS[1] یا خرابی‌های معمول شبکه) می‌تواند مسائل جدی برای سازمان و کاربران که به شبکه به عنوان یک ابزار متکی هستند به وجود آورد[۱].

نفوذها به دو دسته داخلی و خارجی تقسیم می‌شوند. نفوذهای خارجی به نفوذهایی گفته می‌شوند که توسط افراد مجاز یا غیرمجاز، از خارج شبکه به درون شبکه صورت می‌گیرد. نفوذهای داخلی توسط افراد مجاز در سیستم و شبکه داخلی، از درون خود شبکه انجام می‌پذیرد. یک سیستم تشخیص نفوذ تمام فعالیت‌های شبکه را برای شناسایی حملات شناخته شده و یا ناشناخته نظارت می‌کند و برای تشخیص اینکه این وقایع نشانه‌ای از یک حمله یا استفاده مشروع از سیستم هستند، تصمیم می‌گیرد[۲].

مسئله تشخیص نفوذ، فرایند نظارت بر وقایع رخداده در یک شبکه و یا یک سیستم کامپیوتری در جهت کشف موارد انحراف از سیاست‌های امنیتی می‌باشد. سیستم تشخیص نفوذ (IDS[2]) نرم‌افزاری است با قابلیت تشخیص، آشکارسازی و پاسخ (واکنش) به فعالیت‌های غیرمجاز یا ناهنجار که در رابطه با سیستم می‌باشد[۲].

با جمع‌آوری داده‌های ترافیک شبکه و با کمک روش‌های انتخاب ویژگی، تنها بردارهای ویژگی که حاوی اطلاعات ضروری هستند در نظر گرفته می‌شوند. داده‌های جمع‌آوری شده در این مرحله برای تعیین اینکه عادی هستند یا غیرعادی، تجزیه و تحلیل می‌شوند. در این گام از روش‌های مختلف برای تشخیص نفوذ استفاده می‌شود. در انتها واکنش سیستم تشخیص نفوذ است که هشدار لازم را به مدیر سیستم مبنی بر اتفاق افتادن یک حمله و ماهیت آن حمله می‌دهد. همچنین سیستم تشخیص نفوذ نیز در کنترل حملات با بستن پورت شبکه یا از بین بردن فرایندها، کمک می‌کند.

همانطور که پیش‌تر ذکر شد با افزایش مشکلات امنیتی و حملات شبکه در سال‌های اخیر، سیستم‌های تشخیص نفوذ، یک جزء اصلی برای امنیت شبکه به شمار می‌آیند. این سیستم‌ها با داده‌های حجیم برای تحلیل مواجه هستند. بررسی بسیاری از این داده‌ها نشان می‌دهد که بسیاری از ویژگی‌های آنها غیرمفید و بی‌تاثیر هستند. بنابراین حذف ویژگی‌های نامناسب از این مجموعه داده، یک راهکار مناسب برای کاهش مجموعه داده در سیستم‌های تشخیص نفوذ است[۳].

امروزه بیشتر رویکردها در تشخیص نفوذ مربوط به مسئله استخراج ویژگی‌های مهم متمرکز شده است. اما استخراج ویژگی‌ها باعث از دست دادن قسمتی از داده‌ها خواهد شد. بیشتر سیستم‌های تشخیص نفوذ کنونی از تمامی پارامترهای موجود در بسته‌های شبکه برای ارزیابی و کشف الگوهای حملات استفاده می‌نماید، در صورتی که برخی از این پارامترها غیرمرتبط و زائد می‌باشند.

با توجه به اینکه در مجموعه داده‌های سیستم‌های تشخیص نفوذ با تعداد بالایی از ویژگی‌ها رو به رو هستیم، کاهش ابعاد می‌تواند به راحت‌تر شدن تحلیل‌ها، افزایش عملکرد جداکننده، حذف اطلاعات تکراری و غیرمربوط کمک کند.

کاهش حجم داده‌ها در دو شکل امکان‌پذیر است. اول، کاهش تعداد ویژگی‌ها، که منجر به حذف ویژگی‌هایی می‌شود که تاکید کمتری بر سیستم تشخیص نفوذ دارند. دوم، کاهش تعداد نمونه‌ها که منجر به حذف رکوردها و نمونه‌هایی می شود که با حذف آنها، سیستم تشخیص نفوذ با دقت بهتری کار می‌کند. در این مقاله روش‌های کاهش ابعاد و تاثیر آنها در تشخیص نفوذ مورد بررسی و تحقیق قرار گرفته است. هدف این بررسی شناسایی مناسب‌ترین روش کاهش ابعاد در سیستم تشخیص نفوذ براساس ماشین بردار پشتیبان (SVM[3]) است.

ساختار ادامه مقاله به این شرح است: بخش دوم مروری بر روش‌های شناخته شده کاهش ابعاد است. در بخش سوم بررسی تاثیر روش‌های خطی و غیرخطی کاهش ابعاد در سیستم‌های تشخیص نفوذ ارائه خواهد شد. بخش چهارم ارزیابی روش‌های مورد بررسی به همراه جدول نتایج و مقایسه‌ها و تحلیل آنها مطرح می‌شود و نتیجه‌گیری کلی در بخش پنجم بیان شده است.

۲. مروری بر ادبیات پیشین

در این بخش ابتدا بررسی ادبیات موضوع مرتبط با کاهش ابعاد انجام می‌گیرد سپس روش‌های کاهش ابعاد در سیستم های تشخیص نفوذ مطرح می‌شود.

کاهش ابعاد یک راه موثر برای بهبود عملکرد الگوریتم‌های یادگیری ماشین است. کاهش ویژگی‌های بی‌ربط، باعث کاهش زمان آموزش و افزایش دقت طبقه‌بندی می‌شود. انتخاب ویژگی بعنوان یک روش کاهش ابعاد، یک زیرمجموعه ویژگی کاهش یافته را انتخاب می‌کند و همزمان می‌تواند اندازه زیرمجموعه را نیز کاهش دهد و قدرت افتراق را نیز بهبود دهد.

یکی دیگر از جنبه‌های مهم در زمینه‌ی کاهش ابعاد، ارزیابی میزان تاثیرگذاری ویژگی‌های انتخاب شده در طبقه‌بندی می‌باشد. برای ارزیابی میزان تاثیرگذاری ویژگی‌ها، ابتدا ویژگی‌ها را به چندین زیرمجموعه تقسیم کرده، سپس طبقه‌بند را با استفاده از هر یک از این زیرمجموعه ویژگی‌ها، آموزش داده و در نهایت با استفاده از داده‌های آزمون، میزان تاثیرگذاری هر یک از این زیر مجموعه‌ها بدست می‌آید.

الگوریتم‌های مختلفی برای طراحی سیستم‌های تشخیص نفوذ پیشنهاد شده‌اند که دارای قابلیت متعددی هستند. جهت بالا بردن قابلیت یادگیری و کاهش میزان محاسبات در الگوریتم ها از روش‌های کاهش ابعاد استفاده شده است.

یکی از تحقیقات انجام شده، استفاده از تجزیه و تحلیل مولفه اصلی(PCA[۴]) در سیستم تشخیص نفوذ در شبکه است. در[۴]، تاثیر PCA در سیستم تشخیص نفوذ بررسی شده است. همچنین تعداد ایده‌ال مولفه‌های اصلی مورد نیاز برای تشخیص نفوذ و تاثیر داده‌های آلوده به نویز روی PCA نیز مورد توجه بوده است. مجموعه داده‌های با اندازه‌ی اصلی $n \times d$ به ساختاری با k مولفه‌ی اصلی معین، نگاشت شده‌اند و به مجموعه داده‌ای با اندازه‌ی $n \times k$ تبدیل یافته‌اند، که n تعداد نمونه‌ها و d تعداد ابعاد اصلی است. K، تعداد مولفه‌های اصلی با محدوده‌ی تغییرات از ۲ تا ۲۰ است.

نتایج انجام آزمایشات با روش PCA و با استفاده از الگوریتم‌های مختلف طبقه‌بندی مثل $C_{4.5}$، Random Forest، روی دو مجموعه داده به‌نام‌های KDD CUP و UNB ISCX نشان می‌دهد که تعداد ۱۰ مولفه، برای طبقه‌بندی ایده‌ال است. همچنین دقت طبقه‌بندی برای ۱۰ مولفه اصلی به ترتیب در حدود ۹۹,۷ ٪ و ۹۸,۸٪ برای مجموعه داده‌های ذکر شده می‌باشد که تقریبا همان دقت به دست آمده با استفاده از ۴۱ ویژگی از ۳۱۲۷۹ نمونه برای مجموعه داده‌ی KDD و ۲۸ ویژگی از ۳۳۷۴۶ نمونه برای مجموعه داده ISCX می‌باشد.

نتایج این تحقیق نشان داد که نسبت کاهش با PCA برای مجموعه داده KDD Cup و UNB ISCX، به ترتیب ۰,۲۴ و ۰,۳۶ است. همچنین PCA با حضور نویز در داده‌ها، دقت طبقه‌بندی را کاهش داده است و در مقابل هنگامی که داده بدون نویز است افزایش دقت را خواهد داشت. استفاده از PCA برای طراحی یک سیستم تشخیص نفوذ ضمن اینکه پیچیدگی سیستم را کاهش خواهد داد باعث دستیابی به دقت طبقه‌بندی بالاتر نیز می‌شود.

در[۵] یک مدل تشخیص نفوذ با استفاده از ترکیب انتخاب ویژگی Chi-Square و SVM چندکلاسه ارائه شده است. بسیاری از سیستم‌های تشخیص نفوذ، تنها از یک الگوریتم طبقه‌بندی جهت دسته‌بندی ترافیک شبکه بعنوان نرمال و غیرنرمال استفاده می‌کنند. با توجه به میزان زیاد داده‌ها، این مدل از طبقه‌بندها موفق به دستیابی با نرخ تشخیص حمله بالا و کاهش نرخ هشدار غلط نمی‌شوند. با این حال، با استفاده از کاهش ابعاد داده‌ها می‌توانند به یک مجموعه‌ی بهینه از ویژگی‌ها بدون از دست دادن اطلاعات دست یابند و سپس با استفاده از روش مدل‌سازی چندکلاسه، شناسایی حملات شبکه‌ای متفاوت را طبقه‌بندی کنند.

از دیگر تحقیقات انجام شده در[۶] تجزیه و تحلیل ویژگی‌ها، ارزیابی و مقایسه الگوریتم‌های طبقه‌بندی مبتنی بر مجموعه‌ی داده نفوذ دارای نویز می‌باشد. ترافیک داده‌های شبکه در دنیای واقعی با مقدار زیادی از اطلاعات آلوده به نویز همراه است و IDS اغلب در چنین محیطی کار می‌کند. یکی از مسائل چالش برانگیز در IDS برخورد با محیط داده‌ی پر نویز جهت تشخیص حملات از فعالیت‌های شبکه می‌باشد.

در[۶]، الگوریتم‌های داده‌کاوی و یادگیری ماشین مختلف با مجموعه داده‌های NSL-KDD و KDD'99 با نویز ۱۰ درصد و ۲۰ درصد مورد ارزیابی و مقایسه قرار گرفته است. نتایج تجربی نشان می‌دهد که الگوریتم NN(SOM) در مقایسه با دیگر الگوریتم‌های مورد مطالعه از لحاظ مقاوم بودن به محیط نویزی به مراتب بهتر است. با این حال، JRip و J48 از خانواده الگوریتم‌های درخت، عملکرد بهتری نسبت به دیگر الگوریتم‌ها داشتند.

میزان وابستگی ویژگی‌ها در مجموعه داده‌ها برای یک طبقه‌بند خاص توسط روش طبقه‌بندی مبتنی بر عملکرد (PMR°) تجزیه و تحلیل شده است[۶]. ارزیابی نتایج آماری ثابت کرده که هر طبقه‌بند دارای یک ترکیب منحصر به فرد از یک زیرمجموعه ویژگی با نتایج عملکرد مطلوب است و زیرمجموعه‌ی ویژگی انتخاب شده توسط هر الگوریتم طبقه‌بندی، متفاوت از دیگری است. نتایج تجربی نشان داده است که ارزیابی سیستم تشخیص نفوذ براساس مجموعه داده‌ی NSL-KDD دارای نتایج واقعی‌تری در مقایسه با مجموعه داده‌ی اصلی KDD'99 است. نتایج تجربی نشان داده است الگوریتمی که به خوبی روی مجموعه‌ی داده KDD'99 (۱۰٪ داده نویزی) اجرا می‌شود روی مجموعه داده‌ی مشابه NSl-KDD (۲۰٪ داده نویزی) نتیجه نمی‌دهد، لذا این امر ثابت می‌کند مجموعه داده NSL-KDD نشان‌دهنده محیطی واقعی‌تر برای ارزیابی الگوریتم‌های طبقه‌بندی نسبت به مجموعه داده‌ی KDD'99 است[۶].

در[۷]، یک سیستم تشخیص نفوذ برای تشخیص حملات به طور موثر مطرح شده است. برای این منظور، یک الگوریتم انتخاب ویژگی جدید به نام الگوریتم انتخاب ویژگی بهینه مبتنی بر نسبت اطلاعات ارائه شده است. این الگوریتم انتخاب ویژگی، فقط تعدادی از ویژگی‌های مطلوب و مهم را از مجموعه داده‌ی KDD Cup انتخاب می‌کند. علاوه براین، دو روش طبقه‌بندی ماشین بردار پشتیبان و طبقه‌بند مبتنی بر قانون، جهت تاثیر بر طبقه‌بندی داده‌ها و رسیدن به دقت بیشتر استفاده شده است.

در[۸] یکی از چالش‌های مهم در تحقیقات تشخیص نفوذ، طراحی یک سیستم تشخیص نفوذ دقیق از نظر میزان تشخیص بالا، دقت بالا و نرخ هشدار غلط کم است. در این مقاله، یک ساختار کلی از یک رویکرد یادگیری ترکیبی ارائه شده است. سپس روش پیشنهادی با استفاده از خوشه‌بندی K-means و طبقه‌بندهای چندتایی پیاده‌سازی شده است. داده‌ها با روش مبتنی بر الگوریتم خوشه‌بندی K-means تقسیم‌بندی شده است. سپس، هر قسمت با استفاده از یک طبقه‌بند متمایز تقسیم‌بندی شده است. شبکه بیزین، ماشین بردار پشتیبان و الگوریتم‌های طبقه‌بندی OneR بعنوان طبقه‌بند استفاده شده است. روش ترکیبی ارائه شده نتایج بهتری نسبت به تک طبقه‌بند از نظر سرعت کشف و شناسایی، دقت و نرخ هشدار غلط دارا بوده است. نرخ تشخیص روش ترکیبی ارائه شده ۹۹,۵۰٪ است.

در تحقیقات دیگر، یکی از چالش‌های مهم امنیت شبکه، تشخیص ناهنجاری می‌باشد. در[۹] روش جدیدی به نام G-LDA ارائه شد که ترکیب یکپارچه‌سازی تخصیص پنهان دیریکله و الگوریتم ژنتیک با هدف شناسایی ناهنجاری‌ها در ترافیک شبکه بود. علاوه براین، انتخاب ویژگی نقش مهمی را در شناسایی زیرمجموعه‌ی بهینه از ویژگی‌ها برای تعیین بسته‌های ناهنجار ایفا می‌کند. تخصیص پنهان دیریکله، شناسایی مجموعه‌ی بهینه از ویژگی‌ها را برای طبقه‌بندی انجام می‌دهد و از الگوریتم ژنتیک برای محاسبه نمره اقلام داده‌ای و تولید یک جمعیت از زیرمجموعه‌های کاندید استفاده شده است. با استفاده از تابع ارزیابی، میزان شایستگی عناصر جمعیت فعلی مشخص شده و در نهایت بعد از فیلتر شدن عناصر بهتر برای جمعیت نسل بعد انتخاب می‌شوند. این روش بر روی مجموعه داده‌ی KDD-Cup'99 انجام شده و نتایج

تجربی نشان می‌دهد که روش ترکیبی دقت بهتری را برای تشخیص حملات شناخته شده و ناشناخته بدست آورده است. همچنین نرخ مثبت کاذب کمتری نیز گزارش شده است.

از دیگر تحقیقات انجام شده، روش مرکز خوشه و نزدیکترین همسایه است[۱۰]. هدف در سیستم‌های تشخیص نفوذ (IDS) تشخیص انواع مختلفی از ترافیک‌های مخرب شبکه است که توسط یک فایروال عادی و معمولی تشخیص داده نمی‌شود. بسیاری از سیستم‌های تشخیص نفوذ براساس تکنیک‌های یادگیری ماشین توسعه داده شده‌اند. در[۱۰] یک روش نمایش ویژگی جدید به نام روش مرکز خوشه و نزدیکترین همسایه ارائه شده است. روش نمایش ویژگی یک الگوی طبقه‌بند مهم است که کلاس‌بندی صحیح را آسان می‌کند، با این حال، مطالعات کمی روی چگونگی استخراج بیشتر ویژگی‌های مهم برای اتصالات نرمال و تشخیص موثر حملات تمرکز کرده‌اند. در روش مرکز خوشه و نزدیکترین همسایه (CANN[۶]) دو فاصله، اندازه‌گیری می‌شود: فاصله اول مبتنی بر فاصله بین هر نمونه داده و مرکز خوشه‌اش و فاصله دوم بین داده‌ها و نزدیکترین همسایه‌اش در همان خوشه است. نتایج نشان می‌دهد که دقت طبقه‌بندی CANN نسبت به KNN و SVM در مجموعه داده‌ی KDD-Cup’99 بیشتر است. همچنین کارایی محاسباتی بالایی را در زمان آموزش و آزمون طبقه‌بند فراهم می‌آورد.

در تحقیقات دیگر، برای به حداکثر رساندن اثربخشی هر یک از الگوریتم‌های استخراج ویژگی و ایجاد یک سیستم تشخیص نفوذ کارآمد، یک مجموعه‌ای از الگوریتم‌های استخراج ویژگی تحلیل الگوی متمایز خطی (LDA[۷]) و PCA پیاده‌سازی شده است[۱۱]. این روش منجر به نتایج خوبی شده است و دقت بیشتری را در مقایسه با یک روش استخراج ویژگی خاص نشان داده است. هدف، بررسی امکان استفاده از مجموعه ویژگی‌های تولید شده توسط دو روش انعطاف‌پذیر استخراج ویژگی برای تشخیص نفوذ است. بنابراین، حضور گروهی از ویژگی‌های مشابه در هر دو مجموعه ویژگی‌های تولید شده توسط دو روش استخراج ویژگی، اهمیت ویژگی‌های مشترک در مجموعه داده‌ی شبکه را تایید می‌کند. با حداقل نگه داشتن تعداد ویژگی‌ها و بدست آمدن یک مجموعه ایده‌آل از ویژگی‌ها، با پیروی از روش‌های ارائه شده در[۱۱] ، دستیابی به دقت بالا در سیستم تشخیص نفوذ محقق شده است. تعداد کم ویژگی‌ها به این معنی است که طبقه‌بند برای آموزش نیاز به اطلاعات کمتری دارد. نتایج تجربی نشان می‌دهد که کاهش ابعاد می‌تواند میزان تشخیص را افزایش دهد، همچنین مجموع روش‌های استخراج ویژگی، به طور مستقیم عملکرد بهتری را در تشخیص میزان نفوذ نشان داده است.

در پژوهش دیگر[۱۲]، به صورت مصنوعی از روش تحلیل مولفه‌ی اصلی مبتنی بر هسته (KPCA[۸]) و ماشین احتمال Minimax برای شناسایی نفوذ استفاده شده است. ویژگی‌های نفوذ توسط KPCA استخراج شده و با ماشین احتمال Minimax طبقه‌بندی شده است. داده‌های آزمایشی از مجموعه داده‌ی KDD’99 می‌باشند. دراین مقاله، جهت استخراج موثر ویژگی‌ها، یک روش جدید برای تشخیص نفوذ مبتنی بر طبقه‌بند ماشین احتمال Minimax با قابلیت استخراج ویژگی توسط KPCA پیشنهاد شده است. شبیه‌سازی نشان می‌دهد که ماشین احتمال Minimax با استخراج ویژگی KPCA می‌تواند عملکرد تعمیم بهتری را از ماشین بردار پشتیبان و ماشین احتمال Minimax بدون استخراج ویژگی بدست آورد. این آزمایش همچنین نشان داد که روش ارائه شده نیاز به زمان آموزش کمتری دارد. بنابراین، الگوریتم طبقه‌بند ماشین احتمال Minimax[۹] با استخراج ویژگی‌ها توسط KPCA برای تشخیص نفوذ بسیار کارآمدتر می باشد.

۳. بررسی تاثیر روش‌های کاهش ابعاد

این بخش نتایج پیاده‌سازی و آزمایشات روش‌های کاهش ابعاد می‌باشد. در بخش اول مجموعه داده به کار برده شده به همراه اطلاعات مربوط به آن آورده شده است. بخش دوم معیارهای ارزیابی و در بخش سوم انتخاب پارامترهای طبقه‌بندی ارائه شده است. در نهایت نتایج آزمایشات و مقایسه‌ها در بخش سوم صورت گرفته است.

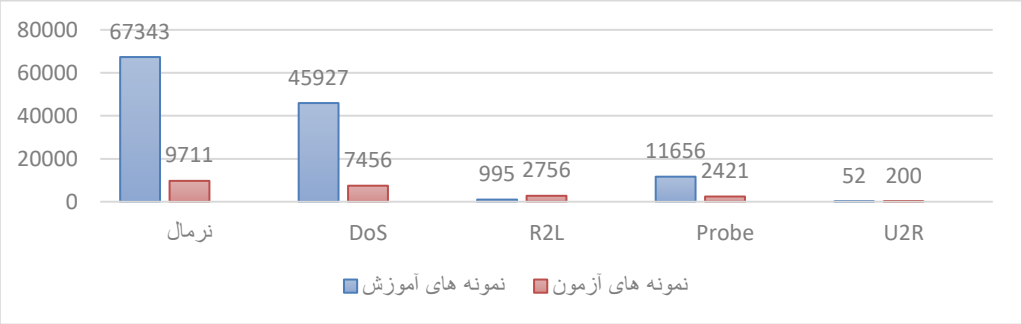
۳- ۱- مجموعه داده

جهت بررسی ارزیابی روش پیشنهادی از مجموعه داده ترافیک شبکه استفاده شده است. این مجموعه داده توسط واحد فناوری و سیستم‌های سایبری آزمایشگاه لینکلن MIT جمع‌آوری شده است. مجموعه داده NSL-KDD (مجموعه داده بهبود یافته KDDcup,1999) با مشخصات مندرج در جدول (۱) در این مقاله استفاده شده است. برای هر اتصال در این مجموعه داده‌ها، ۴۱ ویژگی گسسته و پیوسته تعریف گردیده است و هر اتصال دارای برچسبی است که مشخص می‌کند اتصال مذکور نرمال است یا یکی از انواع حملات. انتخاب مجموعه داده KDD-NSL نسبت به مجموعه داده اصلی KDD این مزیت را دارد که شامل مجموعه رکوردهای تکراری در مجموعه داده یادگیری نمی‌باشد به طوری که طبقه‌بند نسبت به رکوردهای مکرر گرایش پیدا نمی‌کند. این مجموعه داده شامل حملات نوع DoS، R2L، U2R و Probing است.

جدول ۱- اطلاعات مجموعه داده NSL-KDD [۱۳]

تعداد داده‌ها	مجموعه داده‌های آموزشی (KDDtrain+)	مجموعه داده‌های آزمایشی (KDDTest+)
نرمال	67343	9711
حمله	58630	12833
مجموع	125973	22544

به‌طور کلی، حمله‌ها به چهار دسته اصلی تقسیم می‌شوند که عبارتند از: U2R، R2L، DoS و Probe. در شکل ۱ تعداد نمونه‌های آموزش و آزمون در گروه‌های داده‌های نرمال و حمله نشان داده شده است. در این شکل محور افقی نوع داده‌ها و محور عمودی تعداد نمونه‌ها را نشان می‌دهد که در آن داده‌های آزمایش با رنگ آبی و داده‌های آزمون با رنگ قرمز نمایش داده شده‌اند.



شکل ۱- تعداد نمونه‌های آموزش و آزمون در گروه‌های مختلف

۳-۲- معیارهای ارزیابی

ارزیابی داده‌های کاهش یافته حاصل از روش‌های مطرح شده در بخش قبلی توسط طبقه‌بند ماشین بردار پشتیبان صورت می‌گیرد. در این بخش داده‌های کاهش یافته به‌عنوان داده‌های آموزش و آزمون جهت طبقه‌بندی در نظر گرفته می‌شوند که برای این تقسیم‌بندی داده‌ها، از اعتبارسنجی ضربدری ۱۰تایی استفاده شده است. معیارهای ارزیابی در نظر گرفته شده برای بررسی داده‌های تست شامل دقت، صحت و G-mean می‌باشد. معیار G-mean از جمله مهمترین معیارهای ارزیابی داده‌های نامتوازن می‌باشد. با توجه به تفاوت زیاد تعداد نمونه‌های کلاس نرمال با تعداد نمونه‌های هر یک از کلاس‌های حمله، می‌توان این مجموعه داده را به‌صورت نامتوازن در نظر گرفت، لذا معیار G-mean به‌عنوان یک معیار ارزیابی در مساله مورد بررسی، می‌تواند معیاری کاربردی باشد. ابتدا معیارهای ارزیابی را برای کل مجموعه داده با ۴۱ ویژگی، بدون کاهش ابعاد ارزیابی می‌کنیم. نتایج این بررسی در جدول (۲) آورده شده است. صحت تشخیص کلاس عبارتست از تعداد حملات مربوط به کلاس که درست تشخیص داده شده تقسیم بر کل حملات مربوط به آن کلاس، خطای مثبت عبارتست از رفتارهای نرمال که حمله تشخیص داده شده‌اند تقسیم بر کل رکوردهای نرمال.

۳-۳- انتخاب پارامترهای دسته‌بندی

آزمایشات صورت گرفته بر روی روش‌ها توسط نرم‌افزار متلب ۲۰۱۶ بر روی پردازنده GHZ ۴,۲ و ۳۲ گیگا بایت حافظه اصلی انجام شده است. در تمام آزمایشات، از روش اعتبارسنجی ضربدری ۱۰تایی به‌منظور ارزیابی استفاده شده است. در بررسی های انجام شده از هسته تابع پایه‌ای شعاعی استفاده شده است. از طرف دیگر دقت طبقه‌بندی داده‌های کاهش داده شده توسط طبقه‌بند ماشین بردار پشتیبان ارزیابی شده است، به این‌صورت که این طبقه‌بندها با داده‌های کاهش داده شده آموزش می‌بینند سپس با داده‌های آزمون این طبقه‌بندها ارزیابی می‌شوند.

۴. ارزیابی و مقایسه

در این بخش نتایج آزمایشات گزارش شده است. ابتدا، نتایج آزمایش‌ها بدون کاهش ابعاد در مجموعه داده مورد نظر یعنی با ۴۱ ویژگی ارائه شده است، سپس نتایج حاصل از طبقه‌بندی SVM با ورودی‌های تعداد ویژگی‌های متفاوت با ارائه جدول و رسم نمودار تحلیل و بررسی شده است. همچنین نتایج به‌دست آمده از بهترین روش با روش ارائه شده در [۱۴] مقایسه شده است. در جدول (۲) نتایج مربوط به طبقه‌بندی کننده‌های SVM را با ۴۱ ویژگی نشان می‌دهد.

جدول۲- نتایج حاصل از طبقه بندی کننده های SVM با ورودی‌های شامل ۴۱ ویژگی (بدون کاهش ابعاد)

کلاس	صحت	دقت	خطا	میانگین مربعات خطا	G-mean	مدت زمان آموزش(ثانیه)	مدت زمان آزمایش(ثانیه)
Normal	۹۹,۴۸	۹۸,۹۶	۰,۵۱	۰,۵۰	۹۹,۵۰	۱۰۵۲	۳۲,۵۳
Probe	۹۹,۵۷	۹۹,۳۹	۰,۴۲	۰,۴۱	۹۸,۲۲	۹۲۰,۳۹	۳۱,۶۳
DOS	۹۹,۷۴	۹۹,۷۹	۰,۲۵	۰,۲۵	۹۹,۷۱	۱۳۵۵	۳۲,۶۵

با توجه به جدول (۲)، SVM با ۴۱ ویژگی بالاترین دقت طبقه‌بندی را برای کلاس DOS دارد و با توجه به مقادیر دقت و G-mean به‌دست آمده برای این کلاس می‌توان نتیجه گرفت که داده‌های این کلاس با سایر کلاس‌ها تفکیک‌پذیری بیشتری دارد، لذا پیش‌بینی می‌شود که مقادیر دقت و G-mean این کلاس بعد از کاهش نیز زیاد باشد. در جدول (۲)

میانگین زمان آموزش برابر است با ۱۱۰۹,۱۳ ثانیه و میانگین زمان آزمایش برابر با ۳۲,۲۷ ثانیه است. از آنجایی که یکی از اهداف کاهش، بهبود سرعت پردازش داده‌هاست، در جدول (۳) میانگین زمان آموزش و آزمون برای بیشترین (۹ بُعد) و کمترین (۲۳ بُعد) میزان کاهش آورده شده است تا با زمان آموزش و آزمون داده‌های اصلی مقایسه شود.

با توجه به جدول (۳)، کاملا بدیهی است که زمان پردازش داده‌های کاهش یافته کمتر از داده‌های اصلی باشد. در برخی از روش‌ها مثل LDA، kernelPCA، ProbPCA[۱۰] و Autoencoder کاهش زمان پردازش چشمگیری مشاهده می‌شود. لذا با اثبات اینکه کاهش داده‌ها ارتباط مستقیمی با زمان پردازش آنها دارد، در ادامه‌ی آزمایش‌های بعدی به بررسی دقت طبقه‌بندی داده‌های کاهش یافته می‌پردازیم. همانطور که در جدول (۲) ملاحظه شد، SVM با ۴۱ ویژگی، معیار صحت برابر با روش PCA را برای کلاس DOS دارد و بالاترین معیار صحت برای کلاس نرمال مربوط به روش‌های PCA، ProbPCA و MDS[۱۱] می‌باشد. در شناسایی DOS بهترین نتیجه با ویژگی‌های حاصل از FA[۱۲] بدست آمده است.

جدول۳- مقایسه زمان آموزش و آزمایش SVM روش‌های کاهش ابعاد با بیشترین و کمترین میزان کاهش

		نرخ شناسائی با PCA	نرخ شناسائی با LDA	نرخ شناسائی با KPCA	نرخ شناسائی با ProbPCA	نرخ شناسائی با MDS	نرخ شناسائی با Auto Encoder	نرخ شناسائی با FA
۲۳ بعد	میانگین زمان آموزش(ثانیه)	۱۲۶۴	۳۲۱,۷۱	۶۷۱,۰۷	۳۵۷,۵۳	۱۳۰۰	۳۱۵,۶۸	۱۹۶۴
	میانگین زمان آزمایش(ثانیه)	۲۲,۲۹	۳۵,۲۱	۲۴,۵۰	۶,۴۵	۲۲,۲۳	۱۰,۷۷	۳۱,۶۳
۹ بعد	میانگین زمان آموزش(ثانیه)	۶۱۶,۸۲۲	۴۳,۳۲	۹۶,۶۹	۹۲,۹۱۲	۶۱۵,۲۵	۷۳	۷۱۸,۸۸۸
	میانگین زمان آزمایش(ثانیه)	۱۳,۱۴	۴,۳۰	۶,۱۹	۱,۴۷	۱۳,۲۲۲	۶,۰۴	۱۳,۸۶۴

همانطور که گفته شد آزمایش‌ها بر روی داده‌های کاهش یافته با ابعاد ۹، ۱۴، ۱۷، ۲۰ و ۲۳ بُعد، برای سه کلاس شامل یک کلاس نرمال و دو کلاس حمله، برای معیارهای ارزیابی شامل دقت، صحت، میانگین مربعات خطا، زمان اجرا و G-mean صورت گرفته است. گزارش تمامی مقادیر به‌صورت جدول به دلیل حجم بالای نتایج امکان‌پذیر نیست، لذا نتایج برای ۲۳ بُعد به‌صورت جدول، کامل ارائه می‌گردد و برای سایر ابعاد به ارائه نمودار بسنده شده است.

جدول۴- نتایج حاصل از طبقه بندی کننده های SVM با ورودیهای شامل ۲۳ ویژگی با PCA

کلاس	صحت	دقت	خطا	میانگین مربعات خطا	G-mean	مدت زمان آموزش(ثانیه)	مدت زمان آزمایش(ثانیه)

Normal	۹۹,۵۲	۹۹,۰۵	۰,۴۷	۰,۴۶	۹۹,۵۴	۱۱۷۸	۲۲,۳۳
Probe	۹۹,۶۱	۹۹,۰۹	۰,۳۸	۰,۳۸	۹۸,۵۱	۱۰۶۵	۲۲,۰۶
DOS	۹۹,۷۳	۹۹,۷۹	۰,۲۶	۰,۲۶	۹۹,۷۰	۱۵۴۸	۲۲,۵۰

جدول۵- نتایج حاصل از طبقه بندی کننده های SVM با ورودیهای شامل ۲۳ ویژگی با KernelPCA

کلاس	صحت	دقت	خطا	میانگین مربعات خطا	G-mean	مدت زمان آموزش(ثانیه)	مدت زمان آزمایش(ثانیه)
Normal	۵۶,۰۴	۵۵,۹۷	۴۳,۹۵	۴۴,۰۲	۹,۷۳	۹۵۱	۳۶,۸۴
Probe	۹۷,۸۳	۱۰۰	۲,۱۶	۲,۱۶	۳۲,۶۱	۲۵,۲۲	۱,۷۱
DOS	۵۸,۳۳	۱۰۰	۴۱,۶۶	۴۱,۷۳	۶,۳۶	۱۰۳۷	۳۴,۹۶

جدول۶- نتایج حاصل از طبقه بندی کننده های SVM با ورودیهای شامل ۲۳ ویژگی با LDA

کلاس	صحت	دقت	خطا	میانگین مربعات خطا	G-mean	مدت زمان آموزش(ثانیه)	مدت زمان آزمایش(ثانیه)
Normal	۵۱,۷۲	۰	۴۸,۲۷	۴۷,۶۵	۰	۴۷۱,۸۳	۵۲,۳۰
Probe	۸۹,۵۸	۰	۱۰,۴۱	۱۰,۲۸	۰	۱۰۳,۶۱	۹,۸۴
DOS	۵۹,۵۱	۰	۴۰,۴۸	۳۹,۹۷	۰	۳۸۹,۷۰	۴۳,۵۰

جدول۷- نتایج حاصل از طبقه بندی کننده های SVM با ورودیهای شامل ۲۳ ویژگی با ProbPCA

کلاس	صحت	دقت	خطا	میانگین مربعات خطا	G-mean	مدت زمان آموزش(ثانیه)	مدت زمان آزمایش(ثانیه)
Normal	۹۹,۸۶	۹۹,۸۶	۰,۱۳	۰,۱۳	۹۹,۸۷	۳۴۷,۷۷	۶,۸۳
Probe	۹۹,۹۶	۱۰۰	۰,۰۳	۰,۰۳	۹۹,۸۵	۳۷۵,۳۲	۶,۸۴
DOS	۹۹,۹۸	۹۹,۹۷	۰,۰۱	۰,۰۱	۹۹,۹۹	۳۴۹,۵۱	۵,۷۰

جدول۸- نتایج حاصل از طبقه‌بندی کننده‌های SVM با ورودی‌های شامل ۲۳ ویژگی با MDS

کلاس	صحت	دقت	خطا	میانگین مربعات خطا	G-mean	مدت زمان آموزش(ثانیه)	مدت زمان آزمایش(ثانیه)
Normal	۹۹,۴۱	۹۸,۸۰	۰,۵۸	۰,۵۷	۹۹,۴۳	۱۱۶۲	۲۲,۳۲
Probe	۹۹,۵۵	۹۹,۲۹	۰,۴۴	۰,۴۳	۹۸,۱۷	۱۱۸۸	۲۱,۷۵
DOS	۹۹,۷۷	۹۹,۷۹	۰,۲۲	۰,۲۲	۹۹,۷۵	۱۵۵۲	۲۲,۶۴

جدول۹- نتایج حاصل از طبقه بندی کننده‌های SVM با ورودی‌های شامل ۲۳ ویژگی با FactorAnalysis

کلاس	صحت	دقت	خطا	میانگین مربعات خطا	G-mean	مدت زمان آموزش(ثانیه)	مدت زمان آزمایش(ثانیه)
Normal	۹۶,۸۵	۹۳,۸۷	۳,۱۵	۳,۱۰	۹۶,۹۰	۱۸۸۸	۳۱,۸۱
Probe	۹۸,۶۶	۹۹,۴۴	۱,۳۳	۱,۳۱	۹۳,۶۱	۲۰۵۵	۳۱,۲۴
DOS	۹۷,۰۶	۹۹,۶۴	۲,۹۳	۲,۸۹	۹۶,۳۷	۱۹۵۰	۳۱,۸۶

جدول۱۰- نتایج حاصل از طبقه بندی کننده های SVM با ورودیهای شامل ۲۳ ویژگی با Autoencoder

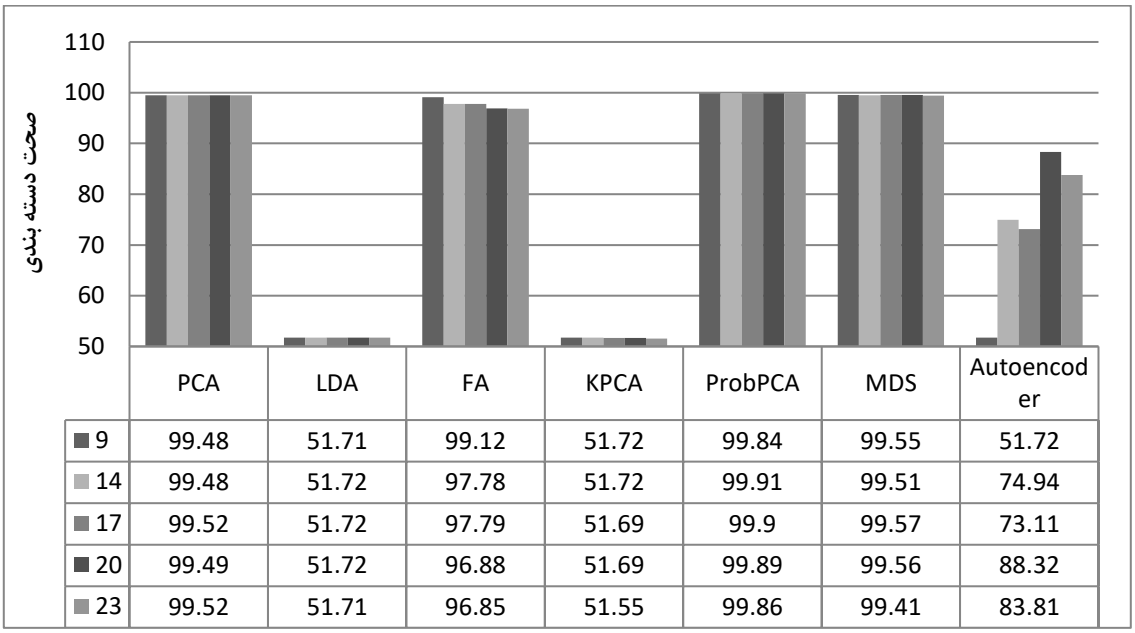
کلاس	صحت	دقت	خطا	میانگین مربعات خطا	G-mean	مدت زمان آموزش(ثانیه)	مدت زمان آزمایش(ثانیه)
Normal	۸۳,۸۱	۸۵,۸۲	۱۶,۱۸	۱۵,۹۷	۸۳,۵۸	۲۲۵,۲۹	۱۵,۶۹
Probe	۸۹,۵۸	۰	۱۰,۴۱	۱۰,۲۸	۰	۳۸۲,۶۷	۶,۳۶
DOS	۹۵,۱۵	۸۹,۵۶	۴,۸۴	۴,۷۸	۹۵,۸۰	۳۳۹,۰۹	۱۰,۲۷

همانطور که ملاحظه می‌شود نتایج حاصل از طبقه‌بندی SVM با ورودی‌های شامل ۲۳ ویژگی با استفاده از روش‌های کاهش ابعاد مختلف و محاسبه معیارهای ارزیابی در جداول (۴) تا (۱۰) گردآوری شده است. از مقایسه سه روش خطی PCA، LDA و FA نسبت به ۴۱ ویژگی چنین برمی‌آید که روش PCA بالاترین دقت طبقه‌بندی را در کلاس Dos دارا می‌باشد. با استفاده از این روش، با وجود اینکه ابعاد مسئله و تعداد ویژگی‌ها کاهش یافته است، دقت تشخیص و طبقه‌بندی کل سیستم نیز کاهش معناداری نیافته است. بنابراین، چنین نتیجه‌گیری می‌شود که روش PCA در کاهش ابعاد می‌تواند یکی از انتخاب‌های مناسب در جهت کاهش تعداد ویژگی‌ها باشد. از طرفی، همانطور که مشاهده می‌شود طبقه‌بندی SVM با ورودی ۲۳ ویژگی در روش LDA قادر به طبقه‌بندی نبوده است و تمامی نمونه‌ها برچسب کلاس

نرمال را دریافت کرده‌اند و به همین دلیل است که دارای معیار صحت کم و معیارهای دقت (Precision) و GMean صفر است، در واقع هیچ حمله‌ای تشخیص داده نشده است ($TP^{۱۲}=0$).

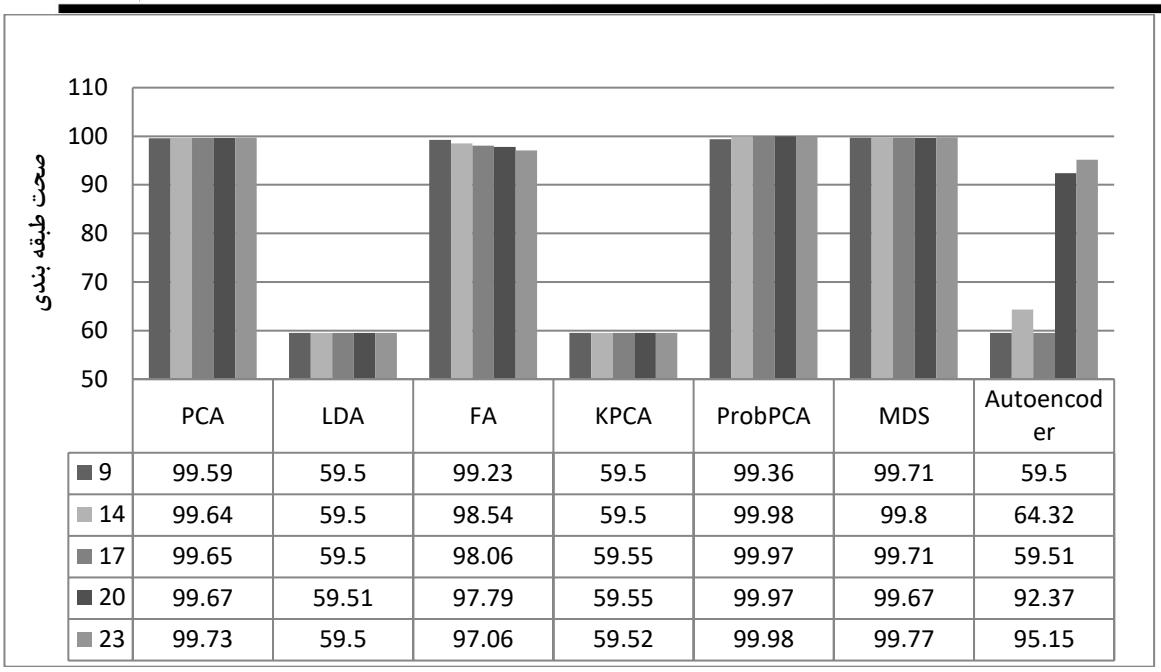
در بررسی روش‌های غیرخطی، با توجه به کاهش شدید معیارهای ارزیابی نظیر دقت و صحت در روش Autoencoder، چنین استنتاج می‌گردد که کاهش ابعاد از طریق این روش کارایی مناسبی ندارد. بنابراین، هرچند کاهش ابعاد با این روش می‌تواند در کاهش حجم محاسبات و افزایش راندمان زمانی سیستم موثر باشد، اما به علت افت شدید دقت سیستم، استفاده از این روش مجاز نیست. همچنین با توجه به نتایج بدست آمده از آزمایشات، ProbPCA با بالاترین دقت و به دلیل احتمالاتی بودن این روش دارای کمترین زمان اجرا بوده و بهترین نتایج حاصل شده است و از آنجاییکه روش ProbPCA سرعت همگرایی کمی دارد بنابراین دقت آن بیشتر است. بعد از ProbPCA، روش مقیاس‌گذاری چندبعدی یا MDS بهترین نتایج را دارا می‌باشد. بالاترین دقت طبقه‌بندی کلاس نرمال در بین دیگر روش‌ها مربوط به روش غیرخطی ProbPCA می‌باشد.

در ادامه معیارهای به‌دست آمده برای سایر ابعاد کاهش یافته بررسی می‌شود. مقادیر به‌دست آمده در آزمایش‌ها برای دو معیار دقت و GMean و همچنین دو معیار خطا و میانگین مربعات خطا بسیار به هم نزدیک هستند. لذا در ادامه معیارهای ارائه شده توسط نمودارهای شکل‌های (۲) تا (۷) برای کلاس‌های مختلف داده‌ها شامل صحت و GMean می‌باشد.



	PCA	LDA	FA	KPCA	ProbPCA	MDS	Autoencoder
9	99.48	51.71	99.12	51.72	99.84	99.55	51.72
14	99.48	51.72	97.78	51.72	99.91	99.51	74.94
17	99.52	51.72	97.79	51.69	99.9	99.57	73.11
20	99.49	51.72	96.88	51.69	99.89	99.56	88.32
23	99.52	51.71	96.85	51.55	99.86	99.41	83.81

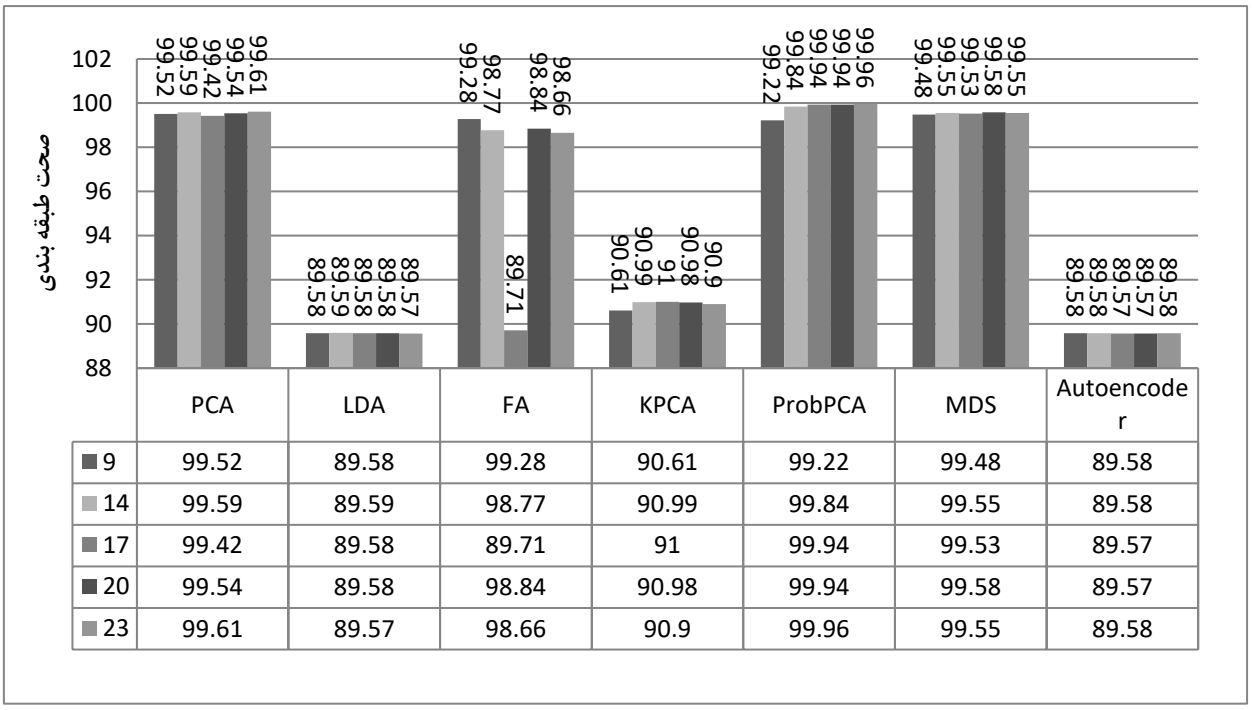
شکل۲- صحت طبقه‌بندی آزمایش روشهای مختلف در کلاس Normal



	PCA	LDA	FA	KPCA	ProbPCA	MDS	Autoencoder
9	99.59	59.5	99.23	59.5	99.36	99.71	59.5
14	99.64	59.5	98.54	59.5	99.98	99.8	64.32
17	99.65	59.5	98.06	59.55	99.97	99.71	59.51
20	99.67	59.51	97.79	59.55	99.97	99.67	92.37
23	99.73	59.5	97.06	59.52	99.98	99.77	95.15

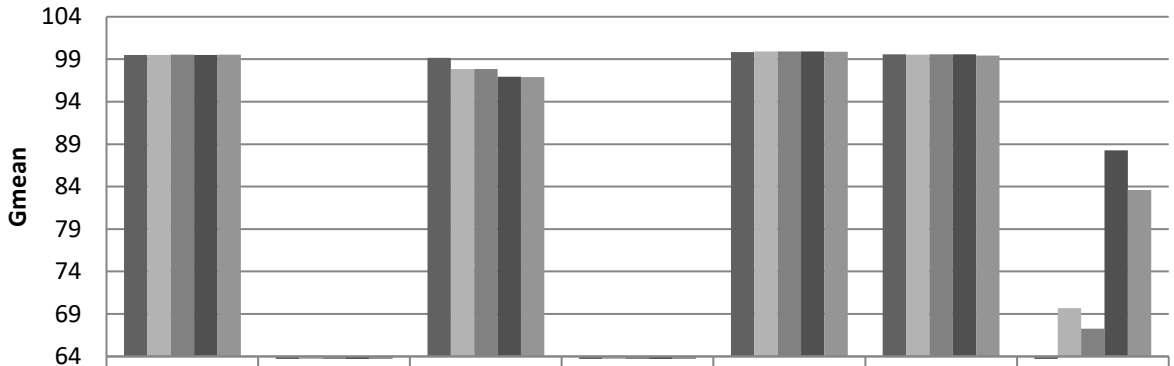
شکل ۳- صحت طبقه‌بندی آزمایش روشهای مختلف در کلاس حمله DOS

-



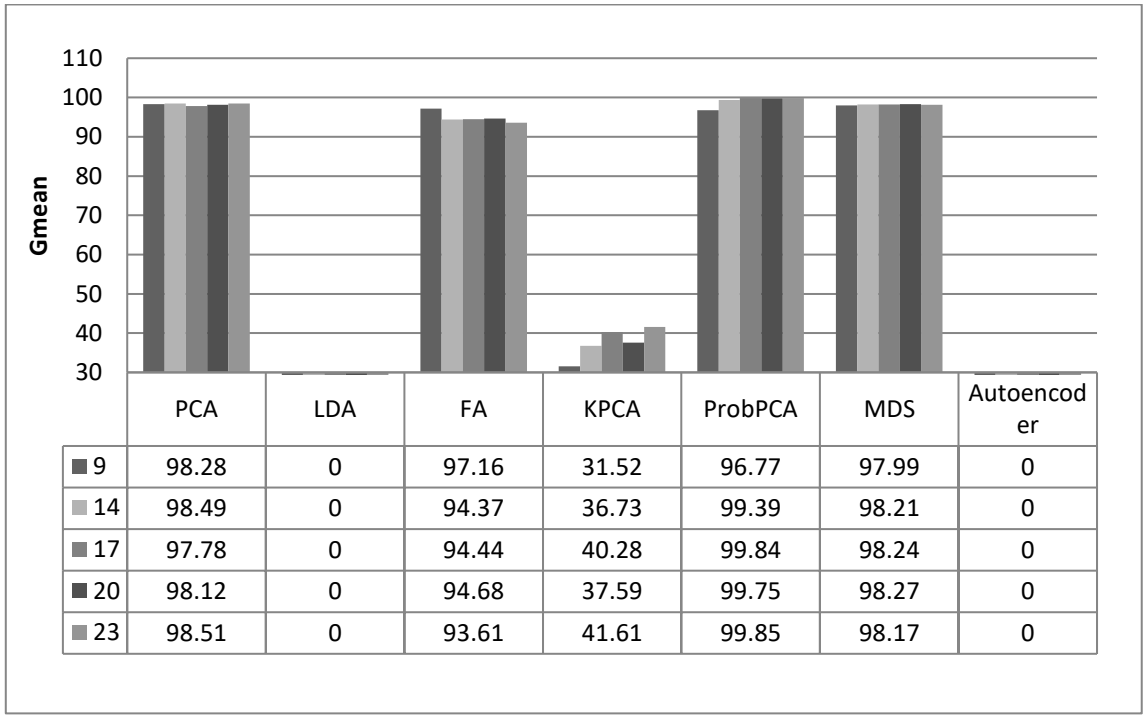
	PCA	LDA	FA	KPCA	ProbPCA	MDS	Autoencoder
9	99.52	89.58	99.28	90.61	99.22	99.48	89.58
14	99.59	89.59	98.77	90.99	99.84	99.55	89.58
17	99.42	89.58	89.71	91	99.94	99.53	89.57
20	99.54	89.58	98.84	90.98	99.94	99.58	89.57
23	99.61	89.57	98.66	90.9	99.96	99.55	89.58

شکل۴- صحت طبقه‌بندی آزمایش روش‌های مختلف در کلاس حمله Probe



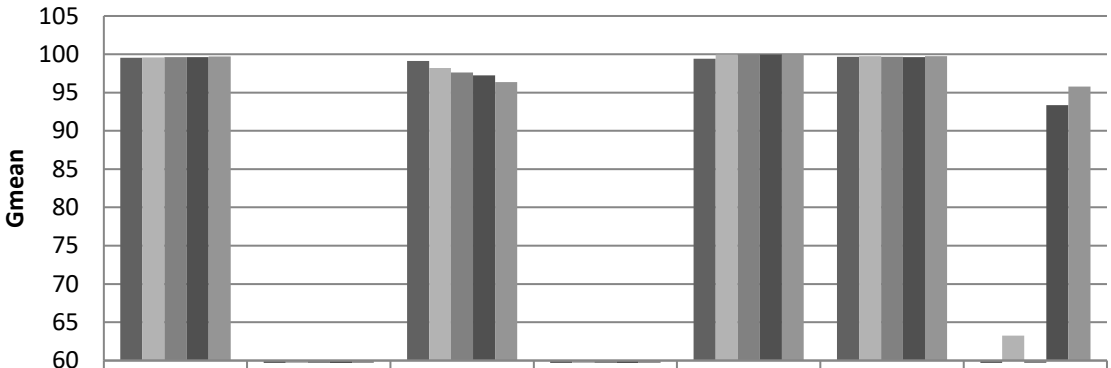
	PCA	LDA	FA	KPCA	ProbPCA	MDS	Autoencoder
9	99.5	0	99.15	0	99.84	99.57	0
14	99.5	0	97.83	0	99.91	99.53	69.67
17	99.54	0	97.84	0	99.91	99.59	67.26
20	99.51	0	96.93	0	99.9	99.58	88.26
23	99.54	0	96.9	0	99.87	99.43	83.58

شکل۵- ارزیابی معیار GMean برای روش‌های مختلف در کلاس حمله Normal



	PCA	LDA	FA	KPCA	ProbPCA	MDS	Autoencoder
9	98.28	0	97.16	31.52	96.77	97.99	0
14	98.49	0	94.37	36.73	99.39	98.21	0
17	97.78	0	94.44	40.28	99.84	98.24	0
20	98.12	0	94.68	37.59	99.75	98.27	0
23	98.51	0	93.61	41.61	99.85	98.17	0

شکل۶- ارزیابی معیار GMean برای روش‌های مختلف در کلاس حمله Probe



	PCA	LDA	FA	KPCA	ProbPCA	MDS	Autoencoder
9	99.55	0	99.11	0	99.43	99.68	0
14	99.6	0	98.22	0	99.98	99.77	63.27
17	99.61	0	97.6	0	99.97	99.68	0
20	99.64	0	97.26	0	99.97	99.63	93.37
23	99.7	0	96.37	0	99.99	99.75	95.8

شکل۷- ارزیابی معیار GMean برای روش‌های مختلف در کلاس حمله DOS

با توجه به نتایج بدست آمده، تغییر میزان ابعاد در روش‌های LDA و MDS بر روی معیارهای ارزیابی تاثیر چشمگیری نداشته است. اما در روش‌هایی مانند PCA، Autoencoder و ProbPCA با تغییر ابعاد می‌توان به جواب بهینه دست یافت که در PCA و Autoencoder با افزایش ابعاد عملکرد بهتری دارد. در روش ProbPCA با کاهش ابعاد برای ۲۰ و ۱۷ بُعد، برای کلاس حمله Probe شاهد کاهش صحت نیستیم اما برای تعداد ۱۴ معیار صحت دچار افت شده است. از سوی دیگر در FA بهترین صحت برای ۹ ویژگی است و هرچه ابعاد افزایش می‌یابد، صحت به مقدار ناچیزی کاهش یافته است. به‌طور کلی، با بررسی معیار صحت برای روش‌های مختلف می‌توان به این نتیجه رسید که تمام روش‌ها به جز روش‌های LDA و Autoencoder (برای کلاس‌های نرمال، Probe و DOS)، صحت بسیار بالایی دارند.

از نظر معیار Gmean، داده‌های کاهش یافته با روش LDA برای تمام ابعاد ضعیف است. می‌توان این‌گونه نتیجه گرفت که روش LDA سبب نامتوازن شدن داده‌ها می‌شود. از طرفی داده‌های کاهش یافته تمام ابعاد با روش‌های PCA، MDS، FA و ProbPCA برای تمام کلاس‌ها از معیار Gmean بالایی برخوردار است. به‌طور کل داده‌های کاهش یافته با روش Autoencoder نتایج قابل قبولی برای Gmean ارائه نداده است. هم‌چنین با توجه به جداول و نمودارها می‌توان نتیجه‌گیری کرد که روش ProbPCA از روش‌های دیگر کاهش بعد مناسب‌تر است.

در جدول (۱۱) نتایج بدست آمده از روش کاهش بعد انتخابی که در مرجع [۱۴] پیاده‌سازی شده و همچنین نتایج حاصل از روش تبدیلی ProbePCA که بهترین روش تبدیل است، آورده شده است. در این جدول، در روش تبدیلی ProbePCA معیار صحت و نرخ تشخیص بهتر از روش انتخابی مرجع [۱۴] است ولی زمان آزمایش بدتر است که علت آن مدت زمانی است که سیستم عملیات محاسباتی برای تبدیل ۴۱ ویژگی به ۲۱ ویژگی کاهش بعد با روش ProbPCA را انجام می‌دهد.

جدول ۱۱- مقایسه کاهش بعد انتخابی در مرجع [۱۴] و روش کاهش بعد تبدیلی در این مقاله

روش کاهش بعد	تعداد مشخصه‌ها	صحت	Normal		Probe		Dos		زمان آزمایش
			FP rate	TP rate	FP rate	TP rate	FP rate	TP rate	
روش ProbPCA پیاده‌سازی شده در این مقاله	۲۱	۹۹,۹۱	۰,۲۷	۹۹,۹۵	۰,۰۵	۹۹,۰۱	۰,۰۱	۹۹,۹۴	۶,۱۳
روش انتخابی در مرجع [۱۴]	۲۱	۹۷,۶۷	۰۱,۷۴	۹۹,۰۷	۰۰,۴۰	۹۰,۳۵	۰۱,۴۹	۹۹,۱۴	۲,۸۴

۵. نتیجه‌گیری

در این پژوهش، کاربردی‌ترین روش‌های خطی و غیرخطی کاهش ابعاد داده‌ها، بر روی مجموعه داده NSL-KDD انجام شد، سپس داده‌های کاهش داده شده حاصل از این روش‌ها، توسط طبقه‌بند ماشین بردار پشتیبان طبقه‌بندی و نتایج حاصل از آن‌ها مورد ارزیابی قرار گرفت. همچنین تاثیر حذف داده‌های پرت نیز بررسی شد، به این‌صورت که جهت دقت طبقه‌بندی داده‌ها ابتدا داده‌های پرت از آن‌ها جدا شده و سپس روش‌های کاهش ابعاد خطی و غیرخطی روی آن‌ها اعمال شده است.

از آنجائیکه بهترین درصدهای شناسایی نفوذ بوسیله سیستم تشخیص نفوذ با ویژگی‌های ProbPCA حاصل شده است، بنابراین می‌توان نتیجه‌گیری کرد که اطلاعات مفید بیشتری در ویژگی‌های بدست آمده از ProbPCA نسبت به روش‌های دیگر وجود دارد. ضمنا با ProbPCA زمان آموزش و آزمایش مناسب‌تری از روش‌های تبدیلی دیگر حاصل می گردد. بنابراین می‌توان نتیجه گرفت که ProbPCA مناسب‌تر از روش‌های دیگر می‌باشد زیرا هم دارای دقت بالاتر و هم زمان محاسباتی قابل قبول‌تر نسبت به سایر روش‌ها است.

در کارهای آینده می‌توان داده‌های کاهش ابعاد یافته در این مقاله (با کمک روش ProbPCA) را با دیگر الگوریتم‌های طبقه‌بندی و یا توسعه‌های بردار پشتیبان نیز مورد آزمایش قرار داد و نتایج طبقه‌بندی را مورد تحلیل و بررسی قرار داد.

۶. مراجع

[1] J. Andress, *THE BASICS OF INFORMATION SECURITY*: Elsevier, 2011.
[2] H. Debar, M. Dacier, and A. Wespi, "Towards a taxonomy of intrusion-detection systems," *Computer Networks*, vol. 31, pp. 805-822, 1999.
[3] S. X. Wu and W. Banzhaf, "The use of computational intelligence in intrusion detection systems: A review," *Applied Soft Computing*, vol. 10, pp. 1-35, 2010.
[4] K. K. Vasan and B. Surendiran, "Dimensionality reduction using Principal Component Analysis for network intrusion detection," *Perspectives in Science*, vol. 8, pp. 510-512, 2016.
[5] I. S. Thaseen and C. A. Kumar, "Intrusion Detection Model using fusion of chi-square feature selection and multi class SVM," *Journal of King Saud University-Computer and Information Sciences*, 2016.

[6] J. Hussain and S. Lalmuanawma, "Feature Analysis, Evaluation and Comparisons of Classification Algorithms Based on Noisy Intrusion Dataset," *Procedia Computer Science*, vol. 92, pp. 188-198, 2016.

[7] S. Balakrishnan, K. Venkatalakshmi, and A. Kannan, "Intrusion detection system using Feature selection and Classification technique," *International Journal of Computer Science and Application*, 2014.

[8] S. V. Farrahi and M. Ahmadzadeh, "KCMC: A Hybrid Learning Approach for Network Intrusion Detection using K-means Clustering and Multiple Classifiers," *International Journal of Computer Applications*, vol. 124, 2015.

[9] B. Kasliwal, S. Bhatia, S. Saini, I. S. Thaseen, and C. A. Kumar, "A hybrid anomaly detection model using G-LDA," in *Advance Computing Conference (IACC), 2014 IEEE International*, 2014, pp. 288-293.

[10] W.-C. Lin, S.-W. Ke, and C.-F. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," *Knowledge-based systems*, vol. 78, pp. 13-21, 2015.

[11] A. A. Aburomman and M. B. I. Reaz, "Ensemble of binary SVM classifiers based on PCA and LDA feature extraction for intrusion detection," in *Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), 2016 IEEE*, 2016, pp. 636-640.

[12] Z. Chen, H. Ren, and X. Du, "Minimax probability machine classifier with feature extraction by kernel PCA for intrusion detection," in *Wireless Communications, Networking and Mobile Computing, 2008. WiCOM'08. 4th International Conference on*, 2008, pp. 1-4.

[13] E. B. M. Tavallaee, W. Lu, and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," *CISDA*, 2009.

[14] R. Singh, H. Kumar, and R. Singla, "An intrusion detection system using network traffic profiling and online sequential extreme learning machine," *Expert Systems with Applications*, vol. 42, pp. 8609-8624, 2015.

---

۱ Denial-Of-service
۲ Intrusion Detection System
۳ Support Vector Machine
٤ Principal Component Analysis
٥ Performance-based Method of Ranking
٦ cluster center and nearest neighbor
۷ Linear Discriminant Analysis
۸ Kernel Principal Component Analysis
۹ Minimax Probability Machine Classifier
۱۰ Probabilistic Principal Component Analysis
۱۱ Multidimensinal Scaling
۱۲ Factor Analysis
۱۳ True Positive