JIPET

Journal of Intelligent Procedures in Electrical Technology
Vol. 16/ No. 62/ Summer 2025
P-ISSN: 2322-3871, E-ISSN: 2345-5594, http://sanad.iau.ir/journal/jipet

Doi: 10.71666/jipet.2024.998705

Research Article

Presenting a New Approach for Detecting Attacks on Voice over Internet Protocol Based on Ensemble Clustering

Farid Bavifard, *Ph.D. Student*, Mohammad Kheyrandish, *Assistant Professor*, Mohammad Mosleh, *Associate Professor*

Department of Computer Engineering- Dezful Branch, Islamic Azad University, Dezful, Iran
bavifard.f@iaud.ac.ir, kheyrandish@iaud.ac.ir, mosleh@iaud.ac.ir

Abstract
Due to lower cost and greater flexibility, voice over internet protocol (VoIP) is widely used in telecommunications. A variety of VoIP terminals causes them to be vulnerable. A common way to secure VoIP includes intrusion detection based on machine learning. Due to the diversity of traffics and lack of class labels for training Intrusion detection systems (IDSs) in many situations, clustering approaches (unsupervised learning) have been focused on. But individual cluster systems can't cover the diversities of feature values well, and some traffic samples may be identified as outliers. As an ensemble approach, the proposed model for solving these problems focuses on using TwoStep clustering algorithm, and by improving it, tries to improve the clustering-based intrusion detection. Moreover, regarding the importance of the feature selection process, a combination of Simulated Annealing algorithm (SA) and Multi-Layer Perceptron (MLP) has been exploited for identifying superior features used for clustering VoIP packets, as Normal or involving DoS, R2L, U2R either Probe attacks. Based on evaluation results obtained on the dataset “Network Security Lab-Knwledge Discovery in Databases” (NSL-KDD) by MATLAB, the proposed feature selection reduced the training and testing times, averagely by 77% and 80%, respectively, by reducing the features to 10 and 8. Also, compared to previous works, the proposed IDS shows average improvements in Accuracy, Detection rate, and F-Measure at 3.34 %, 14.17 %, and 32.87 %, respectively.

Keywords: ensemble clustering, feature selection, intrusion detection system, multi-layer perceptron, optimization algorithm, simulated annealing .

Received: 15 October 2022
Revised: 18 January 2023
Accepted: 11 February 2023

Corresponding Author: Mohammad Kheyrandish

Citation: F. Bavifard, M. Kheyrandish, M. Mosleh, Associate Professor, "Presenting a new approach for detecting attacks on voice over internet protocol based on ensemble clustering", Journal of Intelligent Procedures in Electrical Technology, vol. 16, no. 62, pp. 45-66, June 2025 (in Persian).

Doi: 10.71666/jipet.2024.998705

مقاله پژوهشی

ارائه رویکردی جدید برای تشخیص حملات علیه صدا از طریق پروتکل اینترنت مبتنی بر خوشه‌بندی تجمیعی

فرید باوی‌فرد، دانشجوی دکتری، محمد خیراندیش، استادیار، محمد مصلح، دانشیار

گروه مهندسی کامپیوتر- واحد دزفول، دانشگاه آزاد اسلامی، دزفول، ایران
bavifard.f@iaud.ac.ir, kheyrandish@iaud.ac.ir, mosleh@iaud.ac.ir

چکیده: با توجه به هزینه کمتر و انعطاف‌پذیری بیشتر، انتقال صدا از طریق پروتکل اینترنت (VoIP) به طور گسترده‌ای در ارتباطات راه دور استفاده می‌شود. تنوع پایانه‌های VoIP باعث آسیب‌پذیری آنها می‌شود. یک راه متداول برای ایمن‌سازی VoIP، شامل تشخیص نفوذ مبتنی بر یادگیری ماشین است. با توجه به تنوع ترافیک و عدم وجود برچسب کلاس برای آموزش سیستم‌های تشخیص نفوذ (IDS) در بسیاری از مواقع، بر رویکردهای خوشه‌بندی (یادگیری بدون ناظر) متمرکز شده‌اند. اما سیستم‌های خوشه‌بندی منفرد نمی‌توانند تنوع مقادیر ویژگی‌ها را به‌خوبی پوشش دهند و برخی از نمونه‌های ترافیک ممکن است به‌عنوان نقاط پرت شناسایی شوند. مدل پیشنهادی، به‌عنوان یک رویکرد تجمیعی برای حل این مسائل، روی استفاده از الگوریتم خوشه‌بندی دومرحله‌ای متمرکز شده و سعی می‌کند با ایجاد بهبودی در آن، فرآیند تشخیص نفوذ مبتنی بر خوشه‌بندی را بهبود دهد. علاوه بر این، با توجه به اهمیت فرآیند انتخاب ویژگی، ترکیبی از الگوریتم شبیه‌سازی تبرید (SA) و شبکه عصبی پرسپترون چندلایه (MLP)، برای شناسایی ویژگی‌های برتر مورد استفاده در خوشه‌بندی بسته‌های VoIP، در قالب بسته‌های عادی یا حمله انکار سرویس (DoS)، حمله کاربر به ریشه (U2R)، حمله کاربر از راه دور (R2L) و حمله پویش‌گر مورد بهره‌برداری قرار گرفته است. بر اساس نتایج ارزیابی بر روی مجموعه داده "آزمایشگاه امنیت شبکه– کشف دانش در پایگاه‌های داده‌ای" (NSL-KDD)، توسط نرم‌افزار متلب، انتخاب ویژگی پیشنهادی با کاهش ویژگی‌ها به ۱۰ و ۸، زمان آموزش و آزمایش را به‌ترتیب ۷۷ درصد و ۸۰ درصد کاهش می‌دهد. همچنین در مقایسه با تعدادی از مطالعات قبلی، IDS پیشنهادی بهبود متوسطی معادل ۳/۳۴ درصد، ۱۴/۱۷ درصد و ۳۲/۸۷ درصد را به‌ترتیب در دقت، نرخ تشخیص و معیار F نشان می‌دهد.

کلمات کلیدی: الگوریتم بهینه‌سازی، انتخاب ویژگی، پرسپترون چندلایه، خوشه‌بندی تجمیعی، سیستم تشخیص نفوذ، شبیه‌سازی تبرید .

تاریخ ارسال مقاله: ۱۴۰۱/۷/۲۳
تاریخ بازنگری مقاله: ۱۴۰۱/۱۰/۲۸
تاریخ پذیرش مقاله: ۱۴۰۱/۱۱/۲۲

نام نویسنده‌ی مسئول: محمد خیراندیش
نشانی نویسنده‌ی مسئول: دزفول- بلوار دانشگاه- دانشگاه آزاد اسلامی واحد دزفول- دانشکده فنی و مهندسی

۱- مقدمه

شبکه‌های کامپیوتری هسته اصلی سیستم‌های توزیع اطلاعات را تشکیل می‌دهند و یکی از وقایع مهم در تاریخچه آنها، ابداع روش راه‌گزینی بسته‌ای[۱] بود که از سال ۱۹۷۴، با پیدایش پروتکل کنترل انتقال/پروتکل اینترنت[۲] (TCP/IP)، مورد استفاده گسترده‌تری واقع شد. به دنبال افزایش حجم ارتباطات شبکه‌ای و مطرح شدن مفهوم ترافیک شبکه، کاربردهای مختلفی بر بستر شبکه‌ها شکل یافتند که یکی از مفیدترین آنها، انتقال خطوط تلفن بر روی شبکه و اینترنت، به واسطه فن‌آوری مهم "صدا روی پروتکل اینترنت"[۳] (VoIP) است. در این کاربرد، تماس‌های تلفنی، با استفاده از دروازه‌های[۴] VoIP در یک سو، به‌صورت بسته‌های داده‌ای در می‌آیند و در سوی دیگر، مجدداً بر روی خطوط تلفن قرار می‌گیرند. به واسطه ارزان و فراگیر بودن دسترسی به اینترنت، تسهیل ارتباطات تلفنی و صرفه‌جویی قابل توجه در هزینه‌ها، امروزه وسایل ارتباطی مختلفی، مانند تلفن‌های سخت‌افزاری، گوشی‌های موبایل و حتی کامپیوترهای شخصی، سرویس VoIP را پشتیبانی می‌کنند و امکانات مختلفی نظیر پیام صوتی[۵]، دورنگار[۶]، همایش‌های صوتی و تصویری، ضبط مکالمات، تلفن‌های نرم‌افزاری[۷]، برقراری تماس رایگان با کاربران خارج از سازمان و قابلیت مدیریت متمرکز و گسترش آسان را نیز فراهم می‌سازند.

با وجود همه این مزایا، تنوع پایانه‌های سرویس VoIP سبب می‌شود که این سرویس، از لحاظ امنیتی و در مقابل حملات مختلف آسیب‌پذیر باشد. مهمترین انواع حملات تهدید کننده این سرویس، حمله انکار سرویس[۸] (DoS) و حمله‌های انکار سرویس توزیع شده[۹] (DDoS) هستند؛ شامل ممانعت از دسترسی مجاز به یک سرویس و ارسال درخواست قطع به یک سرویسگر یا تلفن، از طرف یک یا چند میزبان (به‌ترتیب در DoS و DDOS). حملاتی مانند حمله اسمورف یا اسمورفینگ[۱۰]، سیل همگام‌سازی[۱۱] (SYN Flood)، سیل بسته‌های مبتنی بر پروتکل دیتاگرام کاربر[۱۲] (UDP Flood) و سیل بسته‌های مبتنی بر پروتکل پیام کنترل اینترنت[۱۳] (ICMP Flood) از دسته DoS هستند [۱]. از جمله مهمترین حملات DoS، حمله پروتکل پیکربندی میزبان پویا[۱۴] (DHCP)، از طریق ارسال درخواست‌های فراوان به یک سرویسگر DHCP، سرویسگر را تحت فشار قرار می‌دهد تا تمامی آدرس‌های IP ممکن را تخصیص دهد. به‌عنوان نمونه‌ای دیگر، حمله هجوم ناگهانی کاربران[۱۵] (FCA) تعداد بسیار زیادی درخواست ناگهانی را به یک سرویسگر ارسال می‌نماید.

یکی از حملات DDoS، حمله فارمینگ[۱۶]، شامل اتصال به ماشین مشتری به منظور دستیابی به اطلاعات، از طریق یک صفحه وب، پست الکترونیک یا تبادل لحظه‌ای پیام، با استفاده از یک درخواست ظاهراً قانونی است. نوع دیگری از این حمله، بر روی VoIP، شامل انحراف شمار زیادی از مکالمات به یک ناحیه خاص می‌شود.

دسته دیگری از حملات، حملات فرد میانی، شامل خواندن و تغییر پیام‌های مبادله شده بین دو طرف، توسط یک متجاوز و بدون اطلاع آنها است. استراق سمع، جعل و تکرار بسته از این دسته هستند [۲]. حمله دسته‌های ولگرد، شامل انجام اقدامات فریب‌کارانه توسط مهاجمین، جهت دسترسی به تجهیزات و منابع فرد دیگر، نوع دیگری از حملات علیه VoIP است [۲]. همچنین، حمله کلاهبرداری عوارضی[۱۷]، شامل استفاده یک کاربر انتهایی VoIP از سرویسگر VoIP، به منظور برقراری تماس غیرمجاز از طریق شبکه تلفن سوئیچی عمومی[۱۸] (PSTN) سنتی، نوع دیگری از حملات علیه این سرویس است. حمله کد بداندیش نیز از نوع حمله‌های ویروسی است که می‌تواند در برنامه‌های کاربردی مبتنی بر VoIP بسیار مؤثر باشد؛ چراکه تمام این برنامه‌ها دارای آدرس IP بوده و به واسطه اداره مکالمات صوتی، خطر ویروس در آنها افزایش می‌یابد [۳].

اولین گام در راستای مقابله با حملات، جلوگیری از ورود و خروج محتویات غیرمجاز شبکه است که توسط دیواره‌های آتش کنترل می‌شود. از آنجا که عبور عوامل نفوذی از دیواره آتش، علی‌رغم پیشرفت‌های حاصل، اجتناب ناپذیر است، سیستم‌های تشخیص نفوذ و سیستم‌های رمزگذاری، به‌عنوان مکمل امنیتی در کنار دیواره‌های آتش استفاده می‌شوند. سیستم تشخیص نفوذ، دستگاه یا نرم‌افزاری است که ضمن پایش فعالیت‌های شبکه و یا سیستم، وظیفه شناسائی و تشخیص هر گونه استفاده غیرمجاز از منابع شبکه و سوء استفاده، آسیب رسانی و حمله توسط کاربرهای مورد اعتماد و غیرمعتمد، مهاجم‌های بی‌تجربه یا کاربران مجرب را به عهده دارد [۴]. این نوع سیستم‌ها، به طورکلی، شامل دو دسته سیستم‌های مبتنی بر شبکه و مبتنی بر میزبان هستند. دسته اول، عموماً برای پایش[۱۹] ترافیک تمام دستگاه‌ها در شبکه، در نقطه و یا نقاط استراتژیک قرار داده می‌شوند و ضمن مقایسه ترافیک ورودی و خروجی شبکه، با کتابخانه‌ای حاوی حمله‌های شناخته شده، به حملات یا رفتارهای غیرعادی پاسخ

می‌دهند. دسته دوم که تنها بر روی میزبان و یا دستگاه‌های منفرد اجرا می‌شوند، فقط بسته‌های ورودی و خروجی به یک دستگاه را بررسی کرده و در هنگام تشخیص نفوذ و یا فعالیت مشکوک، به مدیر و یا کاربر آن دستگاه هشدار می‌دهند [۴].

از یک چشم‌انداز دیگر، دو روش عمده تشخیص نفوذ، شامل تشخیص موارد سوء استفاده (با تطبیق مدل فعلی شبکه با حملات شناخته شده) و تشخیص ناهنجاری‌ها (با مدل‌سازی رفتار عادی شبکه و تشخیص انحراف از آن) هستند. به دلیل نیاز رویکرد اول به نگهداری پایگاه داده حملات و به صرفه نبودن این کار در اغلب شبکه‌ها، بیشتر از رویکرد دوم استفاده می‌شود و تحقیقات اخیر، به تولید سیستم‌های تشخیص نفوذ هوشمند متمایل شده‌اند [۴].

هوشمندسازی فرآیند شناسایی ناهنجاری‌ها، اغلب متکی بر به‌کارگیری رویکردهای مبتنی بر یادگیری ماشین است؛ یادگیری با ناظر (دسته‌بندی) و یادگیری بدون ناظر (خوشه‌بندی). یکی از مهمترین عوامل تأثیرگذار بر کارایی چنین رویکردهایی، انتخاب مناسب ویژگی‌های نمونه‌های داده‌ای آموزش و آزمایش است؛ چرا که هم از نظر دقت و هم از نظر سرعت، بر کارایی فرآیند تفکیک نمونه‌ها اثر می‌گذارد. این مسئله اغلب به‌عنوان یک مسئله بهینه‌سازی مورد توجه قرار می‌گیرد که توسط الگوریتم‌های تکاملی و فراابتکاری[۲۰] قابل حل است و در این مقاله به‌عنوان یک مسئله فرعی مورد توجه خواهد بود.

روش‌های مبتنی بر خوشه‌بندی، نقاط قوتی مانند توانایی افراز داده‌های فاقد برچسب، قابلیت مقیاس‌پذیری، توانایی مواجهه با انواع داده‌ها، استخراج خوشه‌ها به هر شکل دلخواه، توانایی مقابله با داده‌های نویزی و نادرست و ناقص، عدم حساسیت به ترتیب ورود داده‌ها، عدم نیاز به پارامترهای ورودی و نیز پذیرش داده‌هایی با ابعاد بالا را در حوزه تشخیص نفوذ، از خود نشان داده‌اند و طی دهه‌های گذشته، تعداد زیادی از الگوریتم‌های خوشه‌بندی برای تشخیص نفوذ مورد توجه واقع شده‌اند [۵،۶]. در عین حال، به واسطه ویژگی‌های ذاتی ترافیک‌ها و تنوع رفتارهای حملات، الگوریتم‌های خوشه‌بندی منفرد نتوانسته‌اند به طور مداوم عملکرد خوبی داشته باشند. به علاوه، دستورالعمل روشنی برای انتخاب الگوریتم‌های خوشه‌بندی منفرد وجود ندارد. علاوه بر این، اکثر سیستم‌های خوشه‌بندی سنتی با چالش‌هایی مانند نیاز به فضای ذخیره‌سازی بیشتر برای داده‌های حجیم، افزایش کارایی محاسباتی و کاهش پیچیدگی زمانی نسبتاً بالا و ... سروکار دارند. همچنین، روابط پیچیده و همبستگی قوی بین ویژگی‌های بسته‌های ترافیکی نفوذی باعث می‌شود که یک سیستم خوشه‌بندی نتواند به طور مؤثر بسته‌های ترافیک را از بسته‌های معمولی حاوی حملات مختلف متمایز کند.

یکی از رویکردهای مورد توجه طی سال‌های اخیر، برای جبران ناکارآمدی خوشه‌بندها در انعکاس وابستگی‌های بین ویژگی‌ها، استفاده از سیستم‌های خوشه‌بند تجمیعی[۲۱] بوده است که در آنها، نمونه‌های داده‌ای، توسط چندین خوشه‌بند پایه مستقل خوشه‌بندی شده و خروجی نهایی سیستم، حاصل جمع‌بندی خروجی‌های خوشه‌بندهای پایه است. مزیت چنین سیستمی آن است که ضعف یک خوشه‌بند در توجه به روابط بین ویژگی‌ها (و در نتیجه، خوشه‌بندی نادرست آنها) می‌تواند توسط خوشه‌بند دیگر جبران شود [۷]. یکی از الگوریتم‌های خوشه‌بندی تجمیعی، الگوریتم خوشه‌بندی دو مرحله‌ای[۲۲] است؛ الگوریتمی اکتشافی برای آشکار نمودن خوشه‌های ذاتی و طبیعی موجود در مجموعه داده‌ها که به طور معمول کشف نمی‌شوند. این الگوریتم دارای قابلیت‌هایی نظیر خوشه‌بندی توأم داده‌های کمی و کیفی، انتخاب خودکار تعداد خوشه‌ها (K)، تحلیل کارآمد داده‌های بسیار بزرگ و مشخص کردن مهمترین عوامل مؤثر و مشترک بین خوشه‌ها است. الگوریتم دو مرحله‌ای، با فرض مستقل بودن متغیرها در مدل خوشه‌بندی و در نظر گرفتن اندازه درست‌نمایی، به‌عنوان معیاری برای سنجش فاصله نمونه‌ها، از این درست‌نمایی برای حداکثرسازی مشابهت اعضای هر خوشه استفاده می‌نماید. اگرچه این الگوریتم برای هر داده کمّی، یک توزیع نرمال و برای هر داده کیفی، توزیع چندجمله‌ای را در نظر می‌گیرد، بر اساس آزمون‌های تجربی، نسبت به تغییر فرض‌های استقلال و نرمال بودن، به اندازه کافی قدرتمند[۲۳] است. به این ترتیب، لازمه به‌کارگیری مناسب این الگوریتم تجمیعی برای هدف تشخیص نفوذ، انتخاب مناسب توابع توزیع حاکم بر انواع داده‌ها و پارامترهای آنها و نیز توجه به معیارهای مناسب سنجش شباهت و در وهله دوم، توجه به معیارهای مناسب برای سنجش صحت خوشه‌بندی است. در کنار این چالش‌ها، توجه به رابطه‌های غیرمستقیم بین خوشه‌ها و کشف و نگاشت آنها به مجموعه ویژگی‌های نمونه‌های داده‌ای، به‌عنوان چالش‌های اصلی این مقاله مورد توجه قرار گرفته‌اند.

در راستای حل مسئله انتخاب ویژگی، به واسطه مزایایی نظیر سرعت همگرایی و عدم گرفتاری در بهینه محلی، الگوریتم بهینه‌سازی شبیه‌سازی تبرید[۲۴] (SA)، از میان الگوریتم‌های فراابتکاری، مورد توجه قرار گرفته است و با توجه به نیاز به یک ابزار

جنبی به‌عنوان ارزیاب، از یک شبکه عصبی پرسپترون چندلایه[۲۵] (MLP) استفاده خواهد شد. همچنین، در راستای حل چالش اصلی، یعنی پوشش بیشتر و متنوع‌تر روابط بین نمونه‌ها در فرآیند خوشه‌بندی، الگوریتم خوشه‌بندی دومرحله‌ای به‌گونه‌ای تغییر داده شده که در مرحله اول آن، نمونه‌های داده‌ای، بر اساس روش خوشه‌بندی سلسله مراتبی بالا به پایین، برای تشکیل جمعیت اولیه خوشه‌ها استفاده می‌شود. پیکربندی مناسب خوشه‌بند تجمیعی اصلاح شده، جهت به‌کارگیری در سیستم تشخیص نفوذ VoIP، به‌عنوان نوآوری اصلی این مقاله مورد توجه قرار گرفته است.

در راستای نیل به اهداف ذکر شده، بخش‌های بعدی مقاله به صورت ذیل سازمان‌دهی شده‌اند. بخش دوم، مبانی تحقیق، به معرفی ابزارهای مورد نیاز برای حل مسئله می‌پردازد. در بخش سوم، پیشینه تحقیق، کارهای انجام شده پیشین در حوزه تشخیص نفوذ مرور می‌شوند. بخش چهارم به معرفی رویکرد پیشنهادی اختصاص یافته و در بخش پنجم، روال ارزیابی رویکردهای پیشنهادی معرفی می‌شود و نتایج کلی حاصل از تحقیق در قالب بخش ششم مطرح خواهند شد.

۲- مبانی تحقیق

۲-۱- خوشه‌بندی

روش‌های خوشه‌بندی در یک تقسیم‌بندی، به خوشه‌بندی انحصاری[۲۶]، خوشه‌بندی با همپوشی[۲۷] و خوشه‌بندی سلسله مراتبی[۲۸] تقسیم شده‌اند. نوع اول (مانند روش k میانگین[۲۹])، هر نمونه داده‌ای را دقیقاً به یک خوشه نسبت می‌دهد و نوع دوم (مانند خوشه‌بندی فازی)، به هر نمونه داده‌ای، یک درجه تعلق به ازاء هر خوشه نسبت می‌دهد؛ هر نمونه داده‌ای می‌تواند با نسبت‌های متفاوت، به چندین خوشه تعلق داشته باشد [۸]. در نوع سوم، یک ساختار سلسله مراتبی متشکل از خوشه‌ها تشکیل می‌شود که هر سطح آن، یک خوشه‌بندی مستقل را روی داده‌ها عرضه می‌کند. عملکرد خوشه‌بندی سلسله‌مراتبی معمولاً بر اساس الگوریتم‌های حریصانه[۳۰] و بهینگی مرحله‌ای[۳۱] است و خود به دو شکل بالا به پائین[۳۲] (شروع از مجموعه جامع و انجام تقسیمات تا رسیدن به تعداد خوشه دلخواه) یا پائین به بالا[۳۳] (شروع از خوشه‌های تک عضوی و تجمیع آنها تا رسیدن به سطح مناسب) انجام می‌شود. به‌عنوان الگوریتم‌های خوشه‌بندی سلسله‌مراتبی رایج، می‌توان از الگوریتم‌های پیوند تکی[۳۴]، پیوند کامل[۳۵] و پیوند میانگین[۳۶] نام برد. تفاوت اصلی این روش‌ها به نحوه محاسبه شباهت بین خوشه‌ها مربوط می‌شود [۸].

۲-۱-۱- خوشه‌بندی دو مرحله‌ای

در الگوریتم دومرحله‌ای [۹]، خوشه‌بندی طی دو مرحله انجام می‌شود. در گام اول، با اعمال الگوریتم k میانگین، نمونه‌های داده‌ای یک به یک بررسی شده و بر اساس فاصله با خوشه‌های اولیه تشکیل شده تا آن لحظه و با در نظر گرفتن حد آستانه فاصله قابل قبول، در خوشه‌های موجود ادغام شده و یا خود تشکیل خوشه جدید می‌دهند. از جمله پارامترهای مهم در خوشه‌بندی اولیه، می‌توان به حداکثر تعداد خوشه‌های اولیه، B و آستانه قابل تحمل برای فاصله، T، اشاره نمود. در نتیجه مرحله اول (پیش‌خوشه‌بندی)، مجموعه‌ای از خوشه‌ها تشکیل می‌شوند که هر یک با استفاده از یک سه‌تایی به نام بردار ویژگی خوشه‌بندی (CF) معرفی می‌شود. CF برای خوشه iام که حاوی تمام اطلاعات لازم در مورد خوشه است، به صورت رابطه (۱) قابل تصور است:

$$CF_i = (N_i, LS_i, SS_i) \qquad (۱)$$

که در آن، N_i مبیّن تعداد نمونه‌های داده‌ای عضو خوشه، LS_i حاصل‌جمع خطی تمام نمونه‌های عضو خوشه و SS_i بیانگر حاصل‌جمع خطی مربع‌های ویژگی‌های اعضای خوشه است. در گام دوم، بر اساس روش سلسله مراتبی، تجمیع‌هایی روی خوشه‌های تشکیل شده در مرحله اول صورت می‌گیرد تا خوشه‌ها به تعداد مطلوب برسند. این تعداد می‌تواند بر اساس کاربرد مشخص شده باشد و یا به صورت خودکار معین شود. همانند سایر الگوریتم‌های خوشه‌بندی، فاصله‌ها بر اساس معیار فاصله اقلیدسی (فقط برای داده‌های کمی) یا لگاریتم تابع درست‌نمایی و توزیع مخلوط (برای داده‌های کمی و کیفی) و با استفاده از رابطه‌های (۲) و (۳) محاسبه می‌شوند [۱۰]:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} ((p_i - q_i))^2} \qquad (۲)$$

Table (1): Descriptions of parameters in (4) and (5)

جدول (۱): شرح پارامترهای رابطه‌های (۴) و (۵)

پارامتر	شرح
n_i	تعداد اعضای خوشه iام
n_s	تعداد اعضای خوشه sام
p	تعداد متغیرهای پیوسته
q	تعداد متغیرهای دسته‌بندی شده
m_j	تعداد دسته ها برای متغیر دسته‌ای jام است
$\hat{\sigma}_{ij}^2$	واریانس iامین متغیر پیوسته در بین کل اشیاء داده‌ای
$\hat{\sigma}_i^2$	واریانس jامین متغیر پیوسته برای همه اشیاء داده‌ای در یک خوشه
$\hat{\pi}_{ijl}$	نسبت تعداد اشیاء داده‌ای برای متغیر دسته‌بندی شده در خوشه بر تعداد اشیاء در خوشه

$$d(i,s) = \varepsilon_i + \varepsilon_s - \varepsilon_{(i,s)} \tag{۳}$$

که در آنها، ε_i و ε_s، به‌ترتیب مقادیر پراکندگی خوشه‌های iام و sام و $\varepsilon_{(l,s)}$، فاصله پراکندگی‌های خوشه‌های iام و sام هستند و با استفاده از رابطه‌های (۴)، (۵) و (۶) محاسبه می‌شوند که پارامترهای آنها در جدول (۱) معرفی شده‌اند [۱۰]:

$$\varepsilon_i = n_i \left(\sum_{i=1}^{p} \frac{1}{2} \log\left(\hat{\sigma}_{ij}^2 + \hat{\sigma}_i^2\right) - \sum_{j=1}^{q} \sum_{i=1}^{mj} \hat{\pi}_{ijl} \log\left(\hat{\pi}_{ijl}\right) \right) \tag{۴}$$

$$\varepsilon_s = -n_s \left(\sum_{j=1}^{p} \frac{1}{2} \log\left(\hat{\sigma}_{ij}^2 + \hat{\sigma}_i^2\right) - \sum_{j=1}^{q} \sum_{i=1}^{mj} \hat{\pi}_{ijl} \log\left(\hat{\pi}_{sjl}\right) \right) \tag{۵}$$

$$\varepsilon_{(i,s)} = -n_{(i,s)} \left(\sum_{j=1}^{p} \frac{1}{2} \log\left(\hat{\sigma}_{(i,s)j}^2 + \hat{\sigma}_i^2\right) - \sum_{j=1}^{q} \sum_{i=1}^{mj} \hat{\pi}_{(i,s)jl} \log\left(\hat{\pi}_{(i,s)jl}\right) \right) \tag{۶}$$

یکی از لازمه‌های مهم در روند مرحله دوم، اطمینان از صحت ادغام یا عدم ادغام خوشه‌ها است. برای این منظور، دو معیار مختلف مورد توجه قرار گرفته‌اند: معیار اطلاعات آکائیکه[۳۷] (AIC) و معیار اطلاعات بِیزی[۳۸] (BIC). معیار ارزیابی AIC که حائز توجه بسیاری در مباحث نظریه اطلاعات است، نمایانگر میزان اطلاعاتی است که به واسطه استفاده از روش یا مدل مورد نظر، از دست می‌روند [۱۰]. هر چه مقدار این معیار کوچک‌تر باشد، مدل مورد نظر نسبت به بقیه مدل‌ها، بهتر و مناسب‌تر است. برای محاسبه این معیار، از رابطه (۷) استفاده می‌شود [۱۰]:

$$AIC_k = -2l_k + 2r_k \tag{۷}$$

که در آن، l_k مقدار تابع درست‌نمایی برای خوشه kام است و با استفاده از رابطه (۸) محاسبه می‌شود و r_k تعداد پارامترهای مستقل خوشه k ام، یا به‌عبارتی دیگر، معیار اطلاعات بِیزی (BIC) در مدل توزیع مورد نظر است [۱۰].

$$l_k = \sum_{v=1}^{k} \varepsilon_v \tag{۸}$$

معیار BIC برای ارزیابی مدل‌هایی استفاده می‌شود که بر مبنای استنباط آماری ایجاد می‌شوند. این معیار نیز بیانگر میزان اطلاعات از دست رفته توسط مدل است و ارتباط نزدیکی با معیار ارزیابی AIC دارد؛ با این تفاوت که در محاسبه مقدار BIC، علاوه بر مقدار تابع درست‌نمایی، به تعداد نمونه‌های داده‌ای (مشاهده‌ها) نیز توجه می‌شود. معیار BIC بر اساس رابطه (۹) قابل محاسبه است که در آن، n تعداد نمونه‌های داده‌ای است [۱۰]:

$$BIC_k = -2l_k + r_k \log(n) \tag{۹}$$

۲-۲- انتخاب ویژگی

گاهی به واسطه تنوع و تعداد زیاد ویژگی‌های نمونه‌های داده‌ای، انتخاب مؤثرترین ویژگی‌ها در تفکیک نمونه‌ها از یکدیگر، یک پیش‌نیاز مهم است که اغلب به‌عنوان مسئله انتخاب ویژگی شناخته می‌شود. این مسئله شامل شناسایی ویژگی‌های مهم و حذف

ویژگی‌های فاقد بار اطلاعاتی یا کم تأثیر از بردار ویژگی‌های نمونه‌های داده‌ای است؛ با حفظ دقت و کاهش بار محاسباتی آموزش و عملکرد سیستم یادگیر. با توجه به نوع مسئله، الگوریتم‌های بهینه‌سازی به‌عنوان یکی از بهترین رویکردها برای حل آن مورد توجه قرار گرفته‌اند [۱۱،۱۲]. با وجود روش‌های شمارشی و محاسباتی (جستجوی ریاضی) ، روش‌های ابتکاری[۳۹] و فراابتکاری (جستجوی تصادفی[۴۰]) سازگاری بیشتری از خود نشان داده و مورد محبوبیت بیشتری قرار گرفته‌اند [۱۳].

۱-۲-۲- الگوریتم شبیه‌سازی تبرید

الگوریتم شبیه‌سازی تبرید (SA) که با الهام از فرآیند فیزیکی حداقل‌سازی انرژی درونی فلزات (و کاهش شکنندگی آنها) طراحی شده است، به‌عنوان یک الگوریتم بهینه‌سازی فردگرا شناخته می‌شود. این الگوریتم با یک جواب اولیه تصادفی (جواب بهینه فرضی) و با در نظر گرفتن یک دمای اولیه T شروع شده و سپس در قالب یک حلقه تکرار، به سمت جواب‌های بهتر حرکت می‌کند؛ بهتر شدن جواب‌ها بر اساس یک تابع هدف، E، مشخص می‌شود که متناظر با انرژی درونی جسم در فرآیند فیزیکی است؛ هر جواب جدید در صورتی جایگزین جواب فعلی می‌شود که مقدار تابع هدف برای آن، کمتر از تابع هدف جواب فعلی باشد (ΔE کوچک‌تر یا برابر صفر)؛ در غیر این صورت (اگر $\Delta E>0$ باشد)، جواب جدید بر اساس یک احتمال، جایگزین جواب فعلی می‌شود. این احتمال P، بر اساس رابطه (۱۰) محاسبه می‌شود [۱۴]:

$$p = \exp\left(\frac{-\Delta E}{T}\right) \qquad (۱۰)$$

که در آن، ΔE تفاضل تابع هدف جواب فعلی و جواب همسایه و T پارامتر دما است. چنانچه جواب بهینه طی چند دوره تکرار به یک وضعیت پایدار برسد، دمای T کاهش یافته و فرآیند حرکت به سمت جواب‌های بهینه، در دمای جدید تکرار خواهد شد. شرط خاتمه کل فرآیند آن است که با کاهش پارامتر دما نیز تغییری در جواب بهینه ایجاد نشود.

۳- پیشینه تحقیق

در مرجع [۱۵] یک سیستم تشخیص نفوذ، برای امنیت شبکه، با استفاده از میدان تصادفی شرطی[۴۱] ارائه شد. برای کاهش پیچیدگی، یک روش انتخاب ویژگی با استفاده از الگوریتم oneR پیشنهاد شد. آزمایش‌ها با استفاده از مجموعه داده‌های "آزمایشگاه امنیت شبکه– کشف دانش در پایگاه‌های داده‌ای"[۴۲] (NSL-KDD) انجام و نتایج آزمایش در سیستم پیشنهادی، دقت بالاتری (۹۸ درصد) در شناسایی حمله و همچنین عملکرد بهتری (۹۳ درصد) در شناسایی دسته‌های حمله نشان داد.

در مرجع [۱۶] یک روش ترکیبی جدید برای سیستم تشخیص نفوذ مبتنی بر شبکه ناهنجاری[۴۳] (A-NIDS)، با استفاده از کلونی زنبورهای مصنوعی[۴۴] (ABC) و الگوریتم‌های آدابوست[۴۵]، به منظور دستیابی به نرخ تشخیص[۴۶] (DR) بالا و نرخ مثبت کاذب[۴۷] (FPR) پایین ارائه شد. الگوریتم ABC برای انتخاب ویژگی و ادابوست برای ارزیابی و طبقه‌بندی ویژگی‌ها استفاده شد. شبیه‌سازی بر روی مجموعه داده NSL-KDD انجام شده و روش ترکیبی، عملکرد بهتری را در سناریوهای مختلف مبتنی بر حملات نشان داد. همچنین، دقت و میزان تشخیص این روش در مقایسه با روش‌های دیگر بهبود یافته است.

یک رویکرد یادگیری عمیق برای تشخیص نفوذ، با استفاده از روش هم‌جوشی[۴۸] (ادغام) شبکه عصبی چند کانولوشنی[۴۹] (multi-CNN) در مرجع]۱۷[مطرح شد. نتایج آزمایش که با مجموعه داده NSL-KDD پیاده‌سازی شده است، نشان می‌دهد که روش هم‌جوشی multi-CNN از دقت بالا و پیچیدگی پایینی برخوردار است.

در مرجع]۱۸[، برای حل مشکل راندمان[۵۰] پایین و نرخ مثبت کاذب[۵۱] بالا در تشخیص نفوذ ناشی از افزایش داده‌های با ابعاد بالا، یک الگوریتم ازدحام کریل[۵۲] بهبود یافته بر اساس گام کمند خطی نزدیک‌ترین همسایه[۵۳] (LNNLS-KH) برای انتخاب ویژگی لازم در فرآیند تشخیص نفوذ شبکه ارائه شد. آزمایش‌ها نشان دادند که در مقایسه با الگوریتم‌های CMPSO، ACO، KH و IKH، ویژگی‌ها به اندازه ۴۴ درصد، ۴۲/۸۶ درصد، ۳۴/۸۸ درصد و ۲۴/۳۲درصد نسبت به مجموعه داده NSL-KDD کاهش یافته‌اند. همچنین دقت به میزان ۱۰/۰۳ درصد و ۵/۳۹ درصد افزایش داشته است. الگوریتم LNNLS-KH به سرعت از راه‌حل بهینه محلی خارج می‌شود و عملکرد خوبی را در منحنی تکرار تناسب بهینه، سرعت همگرایی و نرخ تشخیص مثبت کاذب نشان

می‌دهد.

در مرجع [۱۹] یک الگوریتم استخراج ویژگی و تشخیص نفوذ اینترنت اشیا برای شهر هوشمند[۵۴]، مبتنی بر مدل یادگیری مهاجرت عمیق[۵۵] پیشنهاد شد. در بخش آزمایش، از ۱۰ درصد از داده‌ها استفاده شد و نتایج آزمایش نشان می‌دهد که الگوریتم مذکور زمان تشخیص کوتاه‌تر و کارایی تشخیص بالاتری دارد.

در مرجع [۲۰] مدل‌های طبقه‌بندی دو لایه جدیدی بر اساس رویکردهای یادگیری ماشین بِیزی و K نزدیک‌ترین همسایه[۵۶] پیشنهاد شدند. نتایج آزمایش، افزایش مطلوب و امیدوار کننده‌ای را در نرخ تشخیص و هشدار نادرست، در مقایسه با سایر مدل‌های موجود نشان می‌دهد. مدل دو لایه، به دلیل کاهش ابعاد بهینه و انتخاب ویژگی، زمان محاسبات پائینی را فراهم می‌کند. تمام فرآیندهای ارزیابی توسط مجموعه داده‌های NSL-KDD انجام شدند.

در مرجع [۲۱]، برای کاهش مصرف انرژی گره‌ها در شبکه‌های حسگر بی‌سیم[۵۷] (WSN) و در طول پردازش تشخیصی، یک مدل تشخیص نفوذ سلسله‌مراتبی مطرح شد که گره‌ها را در یک WSN، با توجه به عملکردشان خوشه‌بندی می‌کند. نتایج شبیه‌سازی نشان می‌دهد که این سیستم نه تنها دقت تشخیص بالایی را تضمین می‌کند، بلکه زمان تشخیص را نیز به طور چشم‌گیری کاهش می‌دهد.

در مرجع [۲۲] یک رویکرد به نام I-SiamIDS ارائه شد که بهبود یافته Siam-IDS بود. آزمایش‌ها روی مجموعه داده NSL-KDD نشان از برتری I-SiamIDS نسبت به روش‌های دیگر دارد؛ از نظر دقت، بازخوانی، صحت و معیار F[۵۸].

در مرجع [۲۳] برای بهبود عملکرد کلی سیستم تشخیص نفوذ، استفاده از روش‌های یادگیری تنبل[۵۹] پیشنهاد شد و مدل hw-IBK بر اساس مجموعه داده‌های NSL-KDD آموزش داده شده و با دو الگوریتم یادگیری تنبل معروف، IBK و LWL مقایسه شد. نتایج حاصل، الگوریتم‌های تنبل را به‌عنوان راه‌حل پایدار تشخیص نفوذ در شبکه واقعی نشان داده‌اند؛ چرا که ضمن کاهش چشمگیر مدت زمان لازم برای تشخیص تهاجم، تعداد نمونه‌های صحیح طبقه‌بندی شده را بهبود بخشیده است (حدود ۹۷/۵۹ درصد).

در مرجع [۲۴] به بررسی انواع نفوذهای ممکن در سیستم‌ها پرداخته شده و یک مدل سیستم تشخیص نفوذ دو مرحله‌ای پیشنهاد شده است؛ انتخاب ویژگی با استفاده از بهره اطلاعاتی و تشخیص نفوذ، با استفاده از ماشین بردار پشتیبان. نتایج آزمایش مدل مذکور بر اساس مجموعه داده NSL-KDD نشان از برتر بودن آن نسبت به سه الگوریتم خفاش[۶۰]، کلونی زنبور عسل[۶۱] و بهینه‌سازی ازدحام ذرات[۶۲] دارد.

در مرجع [۲۵]، به منظور بهبود دقت تشخیص نفوذ، الگوریتم RBF-SVM را به صورت یک طبقه‌بند آدابوست تنظیم نمودند و با استفاده از تأثیر پارامتر سیگما بر RBF-SVM و ارزیابی تأثیر خطای آموزش بر صافی وزن AdaBoost، یک سیستم تشخیص نفوذ مبتنی بر الگوریتم IABRBFSVM بهبود یافته ارائه نمودند. نتایج تجربی نشان داده‌اند که IABRBFSVM-IDS می‌تواند عملکرد شبکه را بهبود بخشد و آن را به سیستم ایده‌آل نزدیک‌تر کند.

در مرجع [۲۶] رویکرد جدیدی برای طراحی سیستم تشخیص نفوذ، مبتنی بر نمونه‌برداری با استفاده از ماشین بردار پشتیبان، با حداقل مربعات[۶۳] (LS-SVM) ارائه شد و برای نشان دادن اثربخشی روش، آزمایش‌هایی بر اساس پایگاه داده KDD 99 انجام شد. برای ارزیابی عملکرد الگوریتم تشخیص نفوذ مذکور، معیار de facto در نظر گرفته شد و آزمایش‌های انجام شده روی تمام طبقه‌بندی‌های دودویی و چند طبقه‌ای نشان از دقت و کارایی واقع گرایانه آن داشت.

در مرجع [۲۷] با هدف شناسایی نفوذها همراه با کاهش میزان هشدار کاذب و افزایش میزان ردیابی، برای طراحی سیستم‌های تشخیص نفوذ مبتنی بر ناهنجاری، به تکنیک‌های یادگیری ماشین توجه شده است. آزمایش‌ها با استفاده از مجموعه داده NSL-KDD انجام شده و نتایج نشان داده‌اند که استفاده از همبستگی کانونی در انتخاب خصوصیات بهینه شده، دقت FAIS را ۳درصد بهبود می‌بخشد.

در مرجع [۲۸]، برای تشخیص نفوذ در شبکه‌های رایانه‌ای، به جمع‌آوری داده‌های حجیم پرداخته و سه استراتژی انتزاع داده‌ها، یعنی استخراج نمونه، انتخاب ویژگی‌ها و انتزاع ویژگی پیشنهاد شدند. ابتدا یک روش مؤثر برای استخراج نمونه‌های نمایشی از داده‌های عظیم اولیه، قبل از ورود به مدل تشخیص نفوذ، پیشنهاد شد که از دو الگوریتم خوشه‌بندی، انتشار ضریب نفوذ[۶۴] (AP)

و K میانگین سنتی استفاده می‌کرد. سپس از دو استراتژی دیگر، انتخاب ویژگی و استخراج ویژگی در قالب داده‌های انتزاعی، برای تشخیص ناهنجاری استفاده شد. نتایج تجربی نشان داده‌اند که اگرچه هر سه استراتژی راندمان تشخیص را بهبود بخشیده‌اند، استخراج نمونه مبتنی بر AP بهترین عملکرهای انتزاع داده را به‌دست آورده است.

در مرجع [۲۹] یک رویکرد تشخیص ترافیک شبکه‌ای مخرب، با استفاده از شبکه‌های عصبی مصنوعی ارائه شد که مناسب استفاده در سیستم‌های تشخیص نفوذ مبتنی بر بازرسی بسته‌ها بود. نتایج نشان دادند که رویکرد طبقه‌بندی مذکور قادر به شناسایی پوسته کد[۶۵] با دقت بسیار بالا و حداقل تعداد مثبت نادرست، کمتر از ۲ درصد است.

در مرجع [۳۰]، با ارائه مفهوم ماتریس ارتباط متقابل، رابطه وقوع همزمان به‌دست آمد. عیب این ماتریس آن است که فقط رابطه مستقیم وقایع را در نظر می‌گیرد، اما فاقد توانایی در نظر گرفتن اطلاعات غنی از ارتباطات غیرمستقیم در تجمیع است. در واقع، چالش روش مذکور دو جنبه دارد؛ بهبود رابطه شیءگرا با بهره‌گیری از ارتباطات سطح بالاتر (در سطح خوشه) و کشف رابطه ساختار مستقیم و غیرمستقیم در یک مدل یکنواخت.

در مرجع [۳۱]، ارتباطات تجمیعی، با استفاده از گام‌برداری تصادفی[۶۶] روی مجموعه‌ای از قطعات داده‌ای (که به‌عنوان زیرخوشه شناخته می‌شوند) بررسی شدند. اگرچه استفاده از قطعات داده به‌جای نمونه‌های داده‌ای اصلی می‌توانست کارایی محاسباتی بهتری را فراهم کند، اما همچنان دارای دو محدودیت بود؛ افزایش اندازه تجمیع (به‌عنوان مثال، تعداد خوشه‌های پایه)، منجر به افزایش چشمگیر تعداد قطعات تولید شده و در نهایت، افزایش سریع پیچیدگی محاسباتی می‌شود. همچنین، تلاقی دادن چندین خوشه‌بند پایه، منجر به ایجاد وابستگی بین قطعات تولید شده با اندازه‌های بسیار نامتعادل می‌شود.

در مرجع [۳۲]، برای بهبود کارایی خوشه‌بندی تجمیعی، رویکردی مبتنی برگام‌برداری تصادفی، جهت انتشار سریع شباهت‌های خوشه‌گرا ارائه دادند که با کار در سطح شیء [۳۰] و سطح قطعه [۳۱] متفاوت است.

در مرجع [۳۲] کشف اطلاعات غنی تجمیع، در سطح خوشه، با جمع‌بندی[۶۷] چند مقیاسه و نگاشت خوشه-شیء بررسی شدند و نوآوری‌ها، شامل استنتاج یک معیار شباهت خوشه‌گرا (قادر به انعکاس اطلاعات تجمیعی سطح بالاتر)، جادهی ارتباطات غیرمستقیم به وسیله گام‌برداری تصادفی خوشه و ایجاد یک ماتریس همبستگی بهبود یافته[۶۸] (ECA)، بر اساس ترسیم شباهت جدید خوشه‌گرا از شیءهای داده‌ای بود.

در مرجع [۳۳]، روش خوشه‌بندی تجمیعی مبتنی بر پیوند را برای بهبود الگوریتم متا خوشه ارائه نمودند. از مزایای این الگوریتم، در نظرگرفتن اطلاعات مربوط به سطح خوشه و سطح شیء و کشف روابط غیرمستقیم بین خوشه‌ها بود. الگوریتم متاخوشه مبتنی بر پیوند[۶۹] (L-MCLA) از اتصال سه‌تائی وزن‌دار[۷۰] (WCT) برای ساخت ماتریس شباهت خوشه‌ای و کشف روابط غیر مستقیم پنهان استفاده می‌نمود. نتایج آزمایش‌ها نشان دادند که الگوریتم آنها، نه‌تنها نتیجه خوشه‌بندی بهتری ایجاد می‌کند، بلکه کمتر تحت تأثیر اندازه‌های مختلف تجمیع قرار می‌گیرد.

در مرجع [۳۴] روش خوشه‌بندی تجمیعی مبتنی بر پیوند[۷۱] (LCE) برای در نظر گرفتن روابط موجود بین خوشه‌ها ارائه شد. نتایج آزمایش نشان می‌دهد که LCE معمولاً بهتر از الگوریتم‌های خوشه‌بندی تجمیعی موجود در آزمون‌های منفرد عمل می‌کند و با وجود خوشه‌های دارای داده‌های نویزی و نامتعادل، عملکردی عالی و قوی را در انواع مختلف داده ایجاد می‌کند و یک ماتریس سطح بالا ارائه می‌دهد که برای بسیاری از تکنیک‌های خوشه‌بندی عددی قابل استفاده بوده و از نظر محاسباتی نیز، برای مجموعه داده‌های بزرگ کارآمد است.

در مرجع [۳۵] یک انتخاب ویژگی ترکیبی[۷۲] (HFS) همراه با یک طبقه‌بندی تجمیعی پیشنهاد شد که به طور موثر، ویژگی‌های مربوطه را انتخاب و طبقه‌بندی حمله منسجم را ارائه می‌داد. در ابتدا، از نقاط قوت مختلف CfsSubsetEval، جستجوی ژنتیکی و یک موتور مبتنی بر قانون[۷۳]، برای انتخاب مؤثر زیرمجموعه‌های ویژگی‌ها با همبستگی بالا، استفاده می‌کرد که پیچیدگی مدل را به طور قابل توجهی کاهش داده و تعمیم الگوریتم‌های یادگیری را افزایش می‌داد. علاوه بر این، با استفاده از روش رأی‌گیری و میانگین احتمالات، یک طبقه‌بندی تجمیعی ارائه نمود که ترکیبی از الگوریتم K میانگین، ماشین بردار پشتیبان تک کلاسه[۷۴]، خوشه‌بندی فضایی مبتنی بر چگالی کاربردهای نویزی[۷۵] (DBSCAN) و بیشینه‌سازی تابع انتظار[۷۶] (EM)، به اختصار KODE است و به‌عنوان یک طبقه‌بند پیشرفته، به‌طور پیوسته، توزیع‌های احتمال نامتقارن را بین نمونه‌های مخرب و عادی طبقه‌بندی

می‌کرد. در مرجع [۳۶] روش اتصال سه‌تائی وزن‌دار را برای جادهی اطلاعات ماتریس همسایگی مشترک بین خوشه‌ها، در ماتریس ارتباطات همزمان، ارائه نمودند. این ماتریس رابطه مستقیم را پوشش می‌داد؛ اما در پوشش روابط غیرمستقیم ناتوان بود. در مرجع [۳۷] بهره‌برداری از شباهت[۷۷] (SRS) برای پوشش ارتباط غیرمستقیم در ماتریس را مورد توجه قرار دادند؛ اما عیب آن پیچیدگی زمانی بالا و غیرعملیاتی برای مجموعه داده‌های بزرگ بود.

۴- معرفی رویکرد پیشنهادی

در این مقاله، با در نظرگرفتن دو چالش مطرح شده در بخش مقدمه، طراحی یک سیستم تشخیص نفوذ مبتنی بر نسخه بهبود یافته‌ای از الگوریتم خوشه‌بندی دو مرحله‌ای، جهت شناسایی حملات VoIP مورد توجه قرار گرفته است. بر این اساس، رویکرد پیشنهادی در قالب چهار مرحله کلی، در شکل (۱) نشان داده شده است:

- پیش‌پردازش: آماده‌سازی مجموعه دادگان موجود برای فرآیندهای پردازشی بعدی.
- انتخاب ویژگی: انتخاب ویژگی‌های برتر مختص حملات VoIP.
- تشکیل دادگان کاهش یافته: توجه به شاخص‌های ویژگی‌های برتر شناسایی شده و استخراج مقادیر آنها و تشکیل مجموعه دادگانی با نمونه‌های کاهش یافته.
- خوشه‌بندی: تشکیل خوشه‌بند بر اساس رویکرد خوشه‌بندی تجمیعی دو مرحله‌ای بهبودیافته.

۴-۱- مرحله پیش‌پردازش

به دلیل وجود ویژگی‌هایی با ماهیت غیرعددی [ویژگی‌های متنی (اسمی) [۷۸]]، در مجموعه‌های دادگان اولیه، مرحله پیش‌پردازش شامل شناسایی این نوع ویژگی‌ها و تبدیل مقادیر آنها در نمونه‌های داده‌ای، به مقادیر عددی است. برای این منظور، در این مقاله، از یک الگوریتم کدگذاری برای تبدیل این مقادیر به مقادیر عددی استفاده خواهد شد. در این کدگذاری، داده‌های ورودی آموزش و آزمایش از فایل فراخوانی شده و به فرم ماتریسی در می‌آیند و به هر یک از برچسب‌های کاراکتری، یک عدد صحیح منحصر به فرد اختصاص داده می‌شود.

۴-۲- مرحله انتخاب ویژگی

در این مرحله، به واسطه مزیت‌هایی نظیر سرعت همگرایی و عدم گرفتاری در بهینه محلی، الگوریتم بهینه‌سازی شبیه‌سازی تبرید (SA) برای حل مسئله انتخاب ویژگی انتخاب شده و فلوچارت این مرحله در شکل (۲) نشان داده شده است. راه‌حل اولیه در این الگوریتم، مجموعه‌ای تصادفی از شاخص‌های ویژگی‌هاست که به‌عنوان ویژگی‌های برتر اولیه در نظر گرفته می‌شوند. برای ارزش‌گذاری این راه‌حل، از یک شبکه عصبی پرسپترون چندلایه به‌عنوان ارزیاب استفاده می‌شود. شبکه MLP، بر اساس ویژگی‌هایی که شاخص‌های آنها در راه‌حل (قالب ویژگی داوطلب) ذکر شده، آموزش داده شده و آزمایش می‌شود. خطای هر نمونه i توسط رابطه (۱۱) محاسبه می‌شود و متوسط مربع خطای نمونه‌های آزمایش[۷۹] (MSE)، مطابق رابطه (۱۲) محاسبه شده و به‌عنوان خروجی بلوک ارزیاب مورد توجه قرار می‌گیرد:

$$e_i = t_i - y_i \tag{۱۱}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} e_i^2 \tag{۱۲}$$

که در آن، t_i برچسب واقعی و y_i برچسب دسته خروجی تولید شده توسط MLP و N تعداد کل نمونه‌های آزمایشی است. بلوک ارزش‌گذاری، با دریافت مقدار MSE متناظر با قالب ویژگی داوطلب، مقدار تابع هزینه (هدف) متناظر با آن را بر اساس رابطه (۱۳) محاسبه می‌نماید:

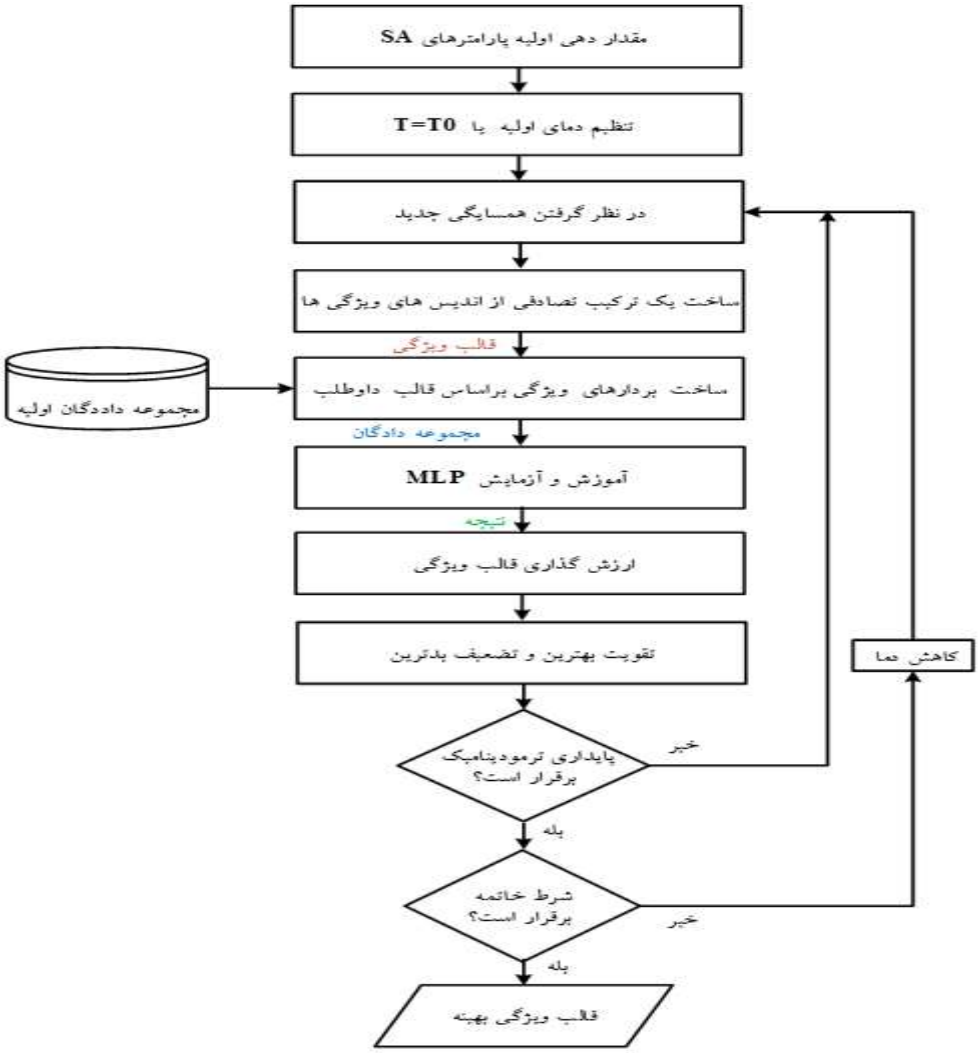
$$E=MSE(1+\beta n_f) \tag{۱۳}$$

که در آن، E نشان‌دهنده تابع هدف الگوریتم SA (متناظر با انرژی)، β هزینه افزودن تعداد ویژگی‌ها و n_f تعداد ویژگی‌های انتخاب شده است.

شکل (۱): ساختار مدل پیشنهادی

Figure (1): The structure of the proposed approach

پس از ارزش‌گذاری راه‌حل اولیه (E_1)، با در نظرگرفتن مقدار اولیه برای پارامتر دما، $T=T_0$، روال تکراری الگوریتم آغاز می‌شود. همان‌طور که در فلوچارت شکل (۲) مشاهده می‌شود، آغازگر این روال، تشکیل یک راه‌حل همسایه جدید، جهت مقایسه با راه‌حل فعلی است. این راه‌حل همسایه، با ایجاد تغییراتی تصادفی در شاخص‌های موجود در راه‌حل فعلی ایجاد می‌شود و دقیقا همانند راه‌حل فعلی، توسط بلوک مبتنی بر MLP، ارزش‌گذاری می‌شود (محاسبه E_2). بر اساس آنچه در بخش ۱-۲-۲ گفته شد، با توجه به مقدار ΔE که از تفاضل E_1 و E_2 به‌دست می‌آید، راه‌حل جدید به‌صورت بلاشرط یا بر اساس احتمال محاسبه شده در رابطه (۹)، جایگزین راه‌حل فعلی خواهد شد و بهترین پاسخ به‌روز رسانی می‌شود. شرط برقراری تعادل ترمودینامیک، عدم تغییر تابع هدف طی ۵ تکرار اخیر در دمای فعلی است. چنانچه این شرط برقرار نباشد، روال گفته شده تکرار خواهد شد. در غیر این‌صورت، شرط خاتمه الگوریتم بررسی می‌شود.



شکل (۲): انتخاب ویژگی بر اساس الگوریتم بهینه‌سازی شبیه‌سازی تبرید

Figure (2): Feature selection based on simulated annealing algorithm (SA)

در صورت برقرار شرط خاتمه و پایان الگوریتم، قالب ویژگی‌های راه‌حل نهایی به عنوان قالب ویژگی برتر اعلام خواهد شد. در غیر این‌صورت (در صورت عدم برقرار شرط خاتمه)، دمان T به اندازه ΔT کاهش یافته و کل روال مجدداً به ازای دمای جدید تکرار خواهد شد. شرط خاتمه مورد توجه در این مرحله، ثابت ماندن راه‌حل برتر طی ۵ دمای اخیر است.

۳-۴- مرحله تشکیل مجموعه دادگان تقلیل یافته

مجموعه دادگان کاهش‌یافته بر اساس قالب ویژگی برتر (حاوی شاخص‌های ویژگی‌های یافته شده در مرحله قبل) تشکیل می‌شود. از میان ویژگی‌های هر نمونه در مجموعه داده‌های اولیه، ویژگی‌های متناظر با شاخص‌های موجود در قالب ویژگی بهینه انتخاب می‌شوند تا نمونه‌ای را در مجموعه داده‌های کاهش یافته تشکیل دهند. هنگام انتقال مقادیر ویژگی‌های انتخاب شده به نمونه جدید ذکر شده، سایر مقادیر ویژگی‌ها کنار گذاشته می‌شوند. بنابراین، مجموعه داده کاهش یافته شامل همان تعداد نمونه‌های مجموعه داده اولیه است؛ اما اندازه هر نمونه برابر با بعد قالب ویژگی برتر است.

۴-۴- مرحله خوشه‌بندی تجمیعی

در این مرحله، انجام یک خوشه‌بندی تجمیعی بر اساس الگوریتم دومرحله‌ای بهبود یافته مورد توجه قرار می‌گیرد. در مرحله اول، پیش‌خوشه‌بندی[۸۰]، نمونه‌های داده‌ای، بر اساس روش خوشه‌بندی سلسله مراتبی بالا به پایین، در قالب مجموعه‌ای از خوشه‌های اولیه تقسیم‌بندی می‌شوند. در این رویکرد، کل نمونه‌های موجود در مجموعه دادگان (خوشه جامع)، به‌تدریج طی چند فاز و بر اساس ویژگی‌های تفکیک‌کننده مؤثر، به زیرمجموعه‌هایی (خوشه‌هایی) با تعداد نمونه‌های کمتر تقسیم می‌شوند و این روال می‌تواند تا رسیدن به خوشه‌های تک عضوی نیز ادامه یابد. در این مرحله، از تابع درست نمائی[۸۱] [رابطه (۲)] به‌عنوان معیار فاصله بین خصوصیت‌های کیفی (با مقادیر گسسته) و فاصله اقلیدسی، برای سنجش فاصله خصوصیت‌های کمی (با مقادیر پیوسته) استفاده می‌شود. در مرحله دوم، زیرخوشه‌های حاصل از مرحله اول، به‌عنوان ورودی‌ها در نظر گرفته شده و بر اساس خوشه‌بندی سلسله مراتبی پایین به بالا، روند ادغام زیرخوشه‌ها تا رسیدن به تعداد مناسبی از خوشه‌ها ادامه می‌یابد. با توجه به این که بسته‌های ترافیکی باید در قالب ۵ نوع عادی[۸۲]، حمله انکار سرویس (DoS)، حمله کاربر از راه دور[۸۳] (R2L)، حمله کاربر به ریشه[۸۴] (U2R) و حمله پویش‌گر[۸۵] تقسیم شوند، لذا تعداد مناسب خوشه‌های نهایی ۵ خواهد بود.

۵- ارزیابی سیستم تشخیص نفوذ پیشنهادی

برای ارزیابی مدل ارائه شده در این مقاله، از نرم‌افزار متلب، جهت پیاده‌سازی مؤلفه‌های تشکیل دهنده استفاده شده است. همچنین، به واسطه مزایایی نظیر عدم وجود رکوردهای تکراری، تعداد مناسب

رکوردهای آموزشی و آزمایشی و قابلیت تشخیص مناسب ناهنجاری‌ها در الگوهای ترافیک شبکه، مجموعه داده NSL-KDD برای آموزش و ارزیابی مدل پیشنهادی مورد توجه قرار گرفته است. در این مجموعه دادگان، هر نمونه داده‌ای شامل ۴۱ ویژگی و یک برچسب نشان دهنده نوع حمله نهفته در بسته ترافیکی متناظر است. ویژگی‌های مذکور، در قالب ویژگی‌های پایه (ویژگی‌های ۱ تا ۹)، ویژگی‌های محتوایی (ویژگی‌های۱۰ تا ۲۲) و ویژگی‌های ترافیکی (ویژگی‌های ۲۳ تا ۴۱) قابل تقسیم هستند و انواع آنها در جدول (۲) طبقه‌بندی شده‌اند.

برچسب‌های نسبت داده شده به هر نمونه، آنها را در قالب دو دسته کلی عادی و غیرعادی تقسیم نموده و برچسب‌های نمونه‌های غیرعادی (۲۲ نوع)، نشان‌دهنده حملات انکار سرویس، کاربر از راه دور، کاربر به ریشه و پویش‌گر در مجموعه داده NSL-KDD هستند. مرجع [۳۸] نمونه‌های مجموعه NSL-KDD را به دو بخش آموزش و آزمایش (به ترتیب $KDDTrain^+$ و $KDDTest^+$) تقسیم‌بندی کرد. تعداد نمونه‌های عادی و غیرعادی (به تفکیک حملات انکار سرویس، کاربر از راه دور، کاربر به ریشه و پویش‌گر) در جدول (۳) و همچنین حملات در مجموعه داده NSL-KDD به تفکیک نوع حمله، در جدول (۴) نشان داده شده است.

Table (2): Types of features in the dataset "Network Security Lab-Knowledge Discovery in Databases" (NSL-KDD) [38]

جدول (۲): نوع ویژگی‌ها در مجموعه داده‌های آزمایشگاه امنیت شبکه– کشف دانش در پایگاه‌های داده‌ای [۳۸]

نوع	ویژگی‌ها
اسمی (متنی)	Protocol_type (2), Service (3), Flag (4)
دودوئی	Land (7), logged_in (12), root_shell (14), su_attempted (15), is_host_login (21), is_guest_login (22)
عددی	Duration (1), src_bytes (5), dst_bytes(6), wrong_fragment(8), urgent(9), hot(10), num_failed_logins(11), num_compromised(13), num_root(16), num_file_creations(17), num_shells(18), num_access_files(19), num_outbound_cmds(20), count(23) srv_count(24), serror_rate(25), srv_serror_rate(26), rerror_rate(27), srv_rerror_rate(28), same_srv_rate(29) diff_srv_rate(30), srv_diff_host_rate(31), dst_host_count(32), dst_host_srv_count(33), dst_host_same_srv_rate(34), dst_host_diff_srv_rate(35), dst_host_same_src_port_rate(36), dst_host_srv_diff_host_rate(37), dst_host_serror_rate(38), dst_host_srv_serror_rate(39), dst_host_rerror_rate(40), dst_host_srv_rerror_rate(41)

Table (3): The number of normal and abnormal samples, in the training and testing subsets [38]

جدول (۳): تعداد نمونه‌های عادی و غیرعادی در مجموعه آموزش و آزمایش [۳۸]

	تعداد کل نمونه‌ها	نرمال	انکار سرویس	کاربر از راه دور	کاربر به ریشه	پویش‌گر
مجموعه آموزش	۱۲۵۹۷۳	۶۷۳۴۳	۴۵۹۲۷	۹۵۵	۵۲	۱۱۶۵۶
مجموعه آزمایش	۲۲۵۴۴	۹۷۱۱	۷۴۵۸	۲۷۵۴	۲۰۰	۲۴۲۱

Table (4): Attacks in NSL-KDD discriminated by attack types [38]

جدول (۴): حملات در مجموعه داده‌های آزمایشگاه امنیت شبکه– کشف دانش در پایگاه‌های داده‌ای به تفکیک نوع حمله [۳۸]

گروه حمله	حملات
انکار سرویس	Back, Land, Neptune, Teardrop, Apache2, Processtable, Udpstorm, POD (Ping of Death), Mailbomb, Fraglle, Smurf
کاربر از راه دور	Guess_Passwd, Imap, Multihop, phf, Warezmastor, SPY, Warezclient, Named, Xlock, FTP_Write, Dictionary, Sendmail, SNMPguess, Imap, Spy, Xsnoop, Httptunnel
کاربر به ریشه	Buffer_Overflow, Loodmodule, Perl, Rootkit, Fdfornat, Xterm, Sqlattack
پویش‌گر	IPsweep, Portsweep, Satan, Nmap, Saint, MScan

مجموعه دادگان NSL-KDD به‌عنوان ورودی فرآیند انتخاب ویژگی مورد توجه قرار گرفته و بر اساس قالب ویژگی برتر یافته شده توسط این فرآیند، مجموعه داده به یک مجموعه داده با نمونه‌های کوچک‌تر کاهش می‌یابد. سپس مجموعه داده کاهش یافته، مجدداً به بخش‌های ذکر شده قبلی تقسیم می‌شود و در فرآیند خوشه‌بندی استفاده می‌شود.

با توجه به دسترسی به برچسب نمونه‌ها در این تحقیق، به منظور ارزیابی سیستم تشخیص نفوذ پیشنهادی، معیارهای خارجی ارزیابی خوشه‌بند مورد توجه قرار گرفته‌اند. برای این منظور، معیارهای تحلیل نتایج، مثبت درست[۸۶] (TP)، منفی درست[۸۷] (TN)، مثبت نادرست[۸۸] (FP) و منفی نادرست[۸۹] (FN) با اعمال داده‌های آزمایش به خوشه‌بند آموزش داده شده، محاسبه می‌شوند و بر اساس آنها، معیارهای جامع‌تر، نظیر دقت[۹۰]، نرخ تشخیص[۹۱]، بازخوانی[۹۲]، صحت[۹۳]، معیار F و نرخ مثبت نادرست[۹۴] با استفاده از رابطه‌های (۱۴) الی (۱۹) ارزیابی می‌شوند [۳۹]:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (۱۴)$$

$$\text{Detection Rate} = \frac{TP}{TP+FP+TN+FN} \quad (۱۵)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (۱۶)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (۱۷)$$

$$\text{F-Measure} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{\text{Precision}+\text{Recall}} \quad (۱۸)$$

$$\text{False Alarm Rate} = \frac{FP}{FP+TN} \quad (۱۹)$$

۵-۱- ارزیابی نقش فرآیند انتخاب ویژگی

در این آزمایش، نحوه تأثیرگذاری فرآیند انتخاب ویژگی بر پارامترهای کارایی سیستم تشخیص نفوذ مورد توجه قرار گرفته است. در جدول (۵)، خوشه‌بندی بدون انجام فرآیند انتخاب ویژگی و خوشه‌بندی بعد از انتخاب ویژگی، از نظر دقت، صحت، بازخوانی و معیارF مقایسه می‌شود.

همچنین، در شکل (۳)، یک مقایسه بین دو حالت فوق (بدون انتخاب ویژگی و با انتخاب ویژگی) از نظر مدت زمان لازم برای آموزش، صورت گرفته است. نتایج نشان داده شده در جدول (۵) و شکل (۳) حاکی از آن است که با وجود تأثیر اندک کاهش ابعاد بر دقت فرآیند خوشه‌بندی، به واسطه انتخاب ویژگی، سرعت فرآیند آموزش خوشه‌بند مبتنی بر ابزارهای مختلف، بهبود قابل ملاحظه‌ای یافته است. علاوه بر این، تأثیر انتخاب ویژگی بر مدت زمان عملکرد سیستم تشخیص نفوذ، در قالب مقایسه انجام شده در جدول (۶)، قابل مشاهده است. این نتایج نشان می‌دهند که کاهش ابعاد نمونه‌های آزمایش، تأثیر قابل ملاحظه‌ای بر سرعت آزمایش و کاهش زمان پردازش دارد.

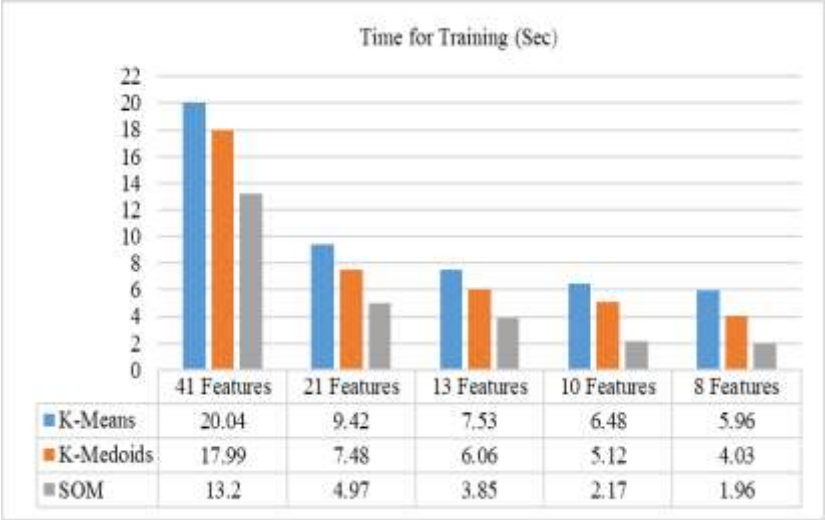
۵-۲- ارزیابی نقش الگوریتم شبیه‌سازی تبرید

در این آزمایش، تأثیر به‌کارگیری الگوریتم شبیه‌سازی تبرید (SA) به‌عنوان الگوریتم بهینه‌سازی جهت انتخاب ویژگی‌های برتر و کاهش ابعاد، مورد ارزیابی قرار می‌گیرد. همان‌طور که نتایج آزمایش در جدول (۷) نشان می‌دهند، الگوریتم شبیه‌سازی تبرید، یک الگوریتم فراابتکاری اثر بخش در حل مسائل بهینه‌سازی در فضاهای جستجوی بزرگ و مجموعه داده‌های با ابعاد بالا است. همچنین، این نتایج نشان می‌دهند که در حل مسائل بهینه‌سازی با استفاده از روش ترکیبی مبتنی بر جایگشت، الگوریتم SA نسبت به سایر الگوریتم‌های فراابتکاری، کارایی و توفیق عمل قابل قبولی کسب نموده است.

Table (5): Comparing clustering before and after feature selection (in terms of Accuracy, Precision, Recall, and F-Measure)

جدول (۵): مقایسه خوشه‌بندی قبل و بعد از انتخاب ویژگی (از نظر دقت، صحت، بازخوانی و معیارF)

روش خوشه بندی	تعداد ویژگی‌ها	دقت (درصد)	صحت (درصد)	بازخوانی (درصد)	معیار F (درصد)
K میانگین	۴۱	۷۷/۲۵	۸۳/۳۹	۷۶/۵۰	۷۹/۸۰
	۲۱	۷۶/۳۲	۸۴/۱۲	۷۷/۲۵	۸۰/۵۳
	۱۳	۷۷/۱۰	۸۳/۸۶	۷۶/۶۱	۸۰/۰۷
	۱۰	۷۶/۸۱	۸۴/۳۷	۷۷/۳۶	۸۰/۷۱
	۸	۷۷/۱۳	۸۴/۹۸	۷۷/۵۵	۸۱/۰۹
K واسط	۴۱	۹۶/۳۷	۹۵/۱۷	۹۶/۴۴	۹۵/۸۰
	۲۱	۹۶/۱۴	۹۵/۶۶	۹۶/۱۵	۹۵/۹۰
	۱۳	۹۶/۰۵	۹۵/۸۹	۹۷/۰۸	۹۶/۴۸
	۱۰	۹۶/۱۹	۹۶/۴۱	۹۷/۲۹	۹۶/۸۴
	۸	۹۵/۹۱	۹۶/۷۳	۹۷/۵۲	۹۷/۱۲
شبکه عصبی خود سازمان‌ده	۴۱	۹۸/۱۸	۹۸/۴۲	۹۸/۱۳	۹۸/۲۷
	۲۱	۹۷/۹۲	۹۸/۱۷	۹۸/۲۹	۹۸/۲۲
	۱۳	۹۷/۶۶	۹۸/۰۲	۹۸/۰۲	۹۸/۰۲
	۱۰	۹۸/۰۴	۹۸/۹۸	۹۸/۸۶	۹۸/۹۱
	۸	۹۷/۳۳	۹۸/۳۵	۹۸/۳۱	۹۸/۳۲



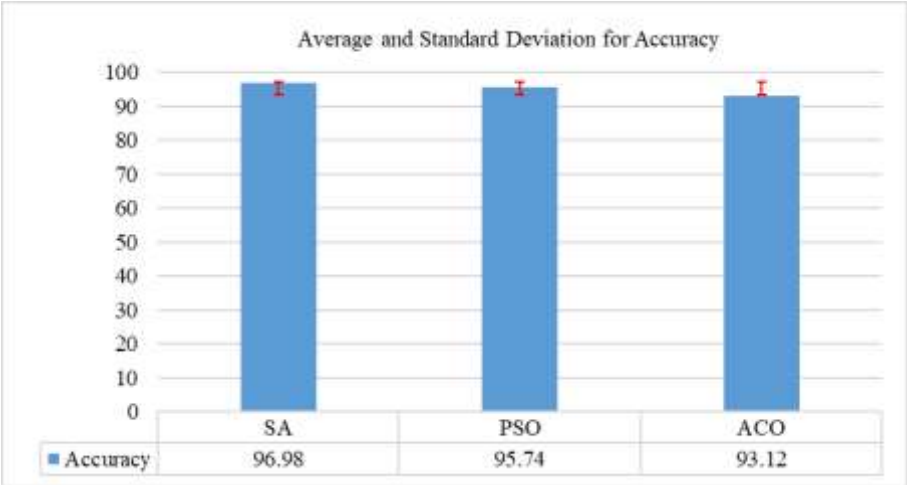
شکل (۳): نموار مقایسه زمان آموزش خوشه‌بندی قبل و بعد از انتخاب ویژگی

Figure (3): Diagram for the comparison on clustering training time before and after feature selection

Table (6): A comparison based on test duration, without feature selection, and after it

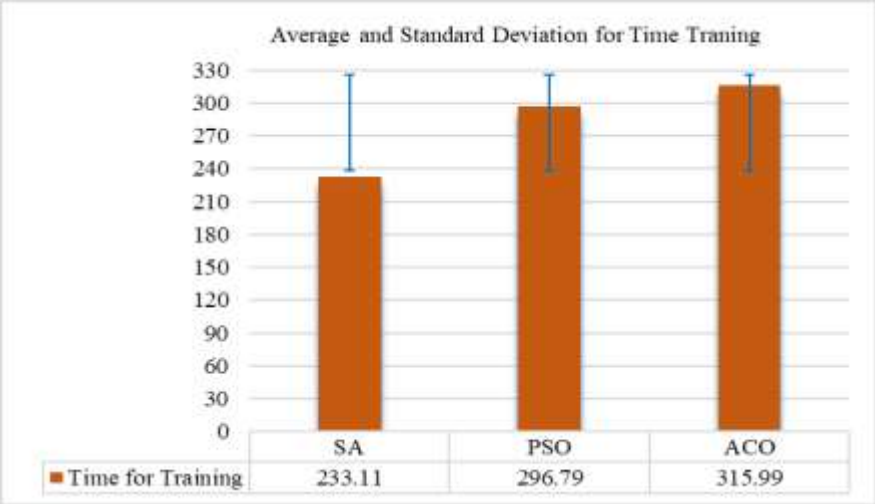
جدول (۶): مقایسه زمان آزمون خوشه‌بندی قبل و بعد از انتخاب ویژگی

تعداد ویژگی‌ها	زمان آزمایش (ثانیه)		
	میانگینK	واسط K	شبکه عصبی خودسازمان‌ده
۴۱	۰/۴۲	۰/۲۱	۰/۱۴
۲۱	۰/۱۶	۰/۰۶	۰/۰۳
۱۳	۰/۱۳	۰/۰۵	۰/۰۲
۱۰	۰/۱۱	۰/۰۴	۰/۰۲
۸	۰/۱۰	۰/۰۴	۰/۰۲



شکل (۴): مقایسه میانگین و انحراف از معیار دقت الگوریتم انتخاب ویژگی پیشنهادی

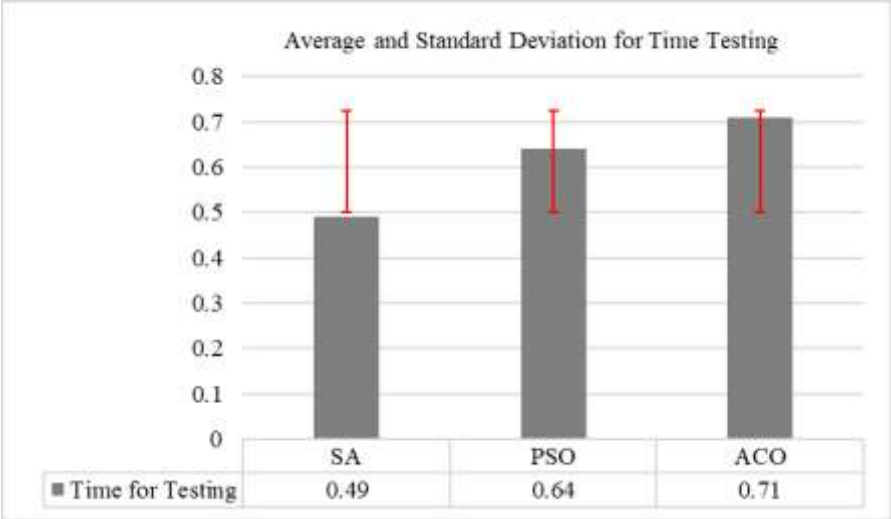
Figure (4): Comparison on the mean and standard deviation of the accuracy criterion, for the proposed feature selection algorithm



شکل (۵): مقایسه میانگین و انحراف از معیار مدت زمان آموزش در الگوریتم انتخاب ویژگی پیشنهادی

Figure (5): Comparison on the mean and standard deviation of training duration for the proposed feature selection algorithm

یکی دیگر از موفقیت‌های نشان داده شده در این آزمایش، سرعت همگرایی و عدم گرفتاری الگوریتم SA در بهینه محلی است که منجر به بهبود راه‌حل نهایی می‌شود. همچنین، نمودار میانگین و انحراف معیار دقت و مدت زمان آموزش و آزمون الگوریتم انتخاب ویژگی پیشنهادی، در مقایسه با دو الگوریتم بهینه‌سازی دیگر، در جدول (۷) و به صورت تفکیک شده، در شکل‌های (۴)، (۵) و (۶) قابل مشاهده است.



شکل (۶): مقایسه میانگین و انحراف از معیار مدت زمان آزمون درالگوریتم انتخاب ویژگی پیشنهادی

Figure (6): Comparison on the mean and standard deviation of test duration for the proposed feature selection algorithm

Table (7): Comparison of optimization algorithms, in terms of feature selection, accuracy and speed

جدول (۷): مقایسه الگوریتم‌های بهینه‌سازی از نظر انتخاب ویژگی، دقت و سرعت

الگوریتم بهینه سازی	ویژگی‌های انتخاب شده	دقت (درصد)	مدت زمان آموزش (ثانیه)	مدت زمان آزمایش (ثانیه)
شبیه‌سازی تبرید	۱۳،۱۶،۲۴،۲۵،۲۸،۳۵،۳۶،۳۷،۳۹،۴۱	۹۶/۸۹(±۰/۹۰)	۲۳۳/۱۱(±۴۵/۴۳)	۰/۴۹(±۰/۱۲)
ازدحام ذرات	۱،۳،۴،۱۳،۱۴،۱۶،۲۱،۲۲،۳۵،۳۶،۳۷،۳۹	۹۵/۷۴(±۱/۳۷)	۲۹۶/۷۹(±۴۶/۵۶)	۰/۶۴(±۰/۱۲)
کلونی مورچگان	۳،۴،۵،۶،۸،۹،۱۰،۱۳،۱۴،۲۵،۳۶،۳۷،۴۱	۹۳/۱۲(±۲/۱۴)	۳۱۵/۹۹(±۴۹/۷۸)	۰/۷۱(±۰/۱۲)

Table (8): Combination of the SA algorithm with the evaluation algorithms and comparing them

جدول (۸): ترکیب الگوریتم شبیه‌سازی تبرید با الگوریتم‌های ارزیاب و مقایسه آنها

ارزیاب	دقت (درصد)	مدت زمان آموزش (ثانیه)	مدت زمان آزمایش (ثانیه)
پرسپترون چندلایه	۹۶/۸۹	۲۳۳/۱۱	۰/۴۹
ماشین بردار پشتیبان	۹۷/۵۳	۲۶۸/۹۷	۰/۶۶
بِیز ساده	۹۶/۴۲	۳۴۲/۸۱	۰/۸۲

این نتایج نشان می‌دهند که الگوریتم انتخاب ویژگی پیشنهادی دارای پراکندگی کمتری در مقادیر معیارهای ارزیابی، در مقایسه با الگوریتم‌های مورد مقایسه است. همچنین بر اساس جدول (۷)، با کاهش تعداد ویژگی‌های نمونه‌های داده‌ای مجموعه NSL-KDD از ۴۱ به ۱۰ ویژگی، کاهش بیشتری را در مقایسه با دو الگوریتم دیگر رقم زده است.

۵-۳- ارزیابی نقش پرسپترون چندلایه در ترکیب با شبیه‌سازی تبرید

در این آزمایش، به‌کارگیری شبکه عصبی پرسپترون چندلایه (MLP)، در مقایسه با ماشین بردار پشتیبان[۹۵] (SVM) و دسته‌بند بِیز ساده[۹۶]، برای تعیین میزان خطای متناظر با قالب بردار ویژگی و محاسبه تابع هدف الگوریتم SA، مورد ارزیابی قرار می‌گیرد. نتایج این آزمایش در قالب جدول (۸) نشان می‌دهند که زمان لازم برای آموزش و آزمایش MLP کوتاه‌تر از الگوریتم‌های SVM و بِیز است و MLP بسیار سریع‌تر از روش‌های مذکور، مرحله یادگیری را پشت سر می‌گذارد. این حکایت از برتری نتایج حاصل از مدل ترکیبی شبکه پرسپترون چندلایه و الگوریتم شبیه‌سازی تبرید دارد.

۵-۴- ارزیابی تأثیر خوشه‌بندی

در این آزمایش، فلسفه استفاده از خوشه‌بندی مورد توجه قرار گرفته و ضمن انجام آزمایش، برتری نسبی خوشه‌بندی بر دسته‌بندی، به‌عنوان فرضیه آزمایش بررسی شده است. نتایج این آزمایش در جدول (۹) گزارش شده‌اند و حکایت از آن دارند که علاوه بر برتری نسبی رویکرد خوشه‌بندی، از نظر معیار دقت ، این رویکرد از نظر سرعت فرآیند یادگیری و پردازش (آزمایش) نیز برتر از رویکرد دسته‌بندی عمل کرده است. علاوه بر این، رویکرد خوشه‌بندی، در قبال نمونه‌های ناشناخته نیز عملکرد بهتری از خود نشان داده است.

۵-۵- ارزیابی تأثیر الگوریتم دومرحله‌ای

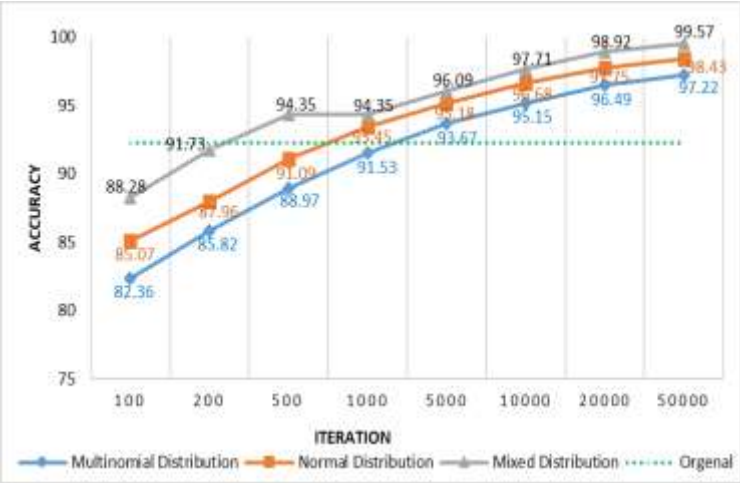
در این آزمایش، به‌کارگیری الگوریتم خوشه‌بندی تجمیعی دومرحله‌ای بهبود یافته در فرآیند تشخیص نفوذ، در مقایسه با الگوریتم‌های خوشه‌بندی منفرد، مورد ارزیابی قرار گرفته است. در این راستا، با در نظر گرفتن چهار نوع حمله انکار سرویس، کاربر از راه دور، کاربر به ریشه و پویش‌گر و حالت عدم حمله (نمونه‌های عادی)، آموزش و آزمایش خوشه‌بند دومرحله‌ای پیشنهادی انجام شده و طی آموزش و آزمایش چند خوشه‌بند منفرد شناخته شده، نتایج حاصل در جدول (۱۰) گزارش شده‌اند.

۵-۶- بررسی تأثیر تابع توزیع

به واسطه اهمیت نقش توابع توزیع در الگوریتم دومرحله‌ای، با انجام آزمایشی، تأثیر توزیع‌های احتمال نرمال[۹۷]، چندجمله‌ای[۹۸] و مخلوط[۹۹] روی پارامترهای دقتی ارزیابی شد. در این آزمایش بر روی ۵ گروه از نمونه‌ها، یعنی نمونه‌های عادی و نمونه‌های حاوی ۴ حمله مختلف انکار سرویس، کاربر از راه دور، کاربر به ریشه و پویش‌گر، اثر تعداد تکرار الگوریتم بر معیارهای کارایی دقت و معیار F مورد توجه قرار گرفت. برای این منظور، مقادیر معیارهای مذکور، به‌طور تفکیک شده برای ۵ نوع ترافیک، پس از ۱۰۰، ۲۰۰، ۵۰۰، ۱۰۰۰، ۱۰۰۰۰، ۲۰۰۰۰ و ۵۰۰۰۰ تکرار الگوریتم ثبت شده و میانگین مقادیر حاصل برای ۵ نوع ترافیک محاسبه شدند. نمودارهای نشان داده شده در شکل‌های (۷) و (۸) حاکی از برتری توزیع مخلوط نسبت به دو توزیع نرمال و چندجمله‌ای است.

۷-۵- مقایسه

در این آزمایش، با رویکرد مقایسه الگوریتم پیشنهادی مبتنی بر خوشه‌بندی دومرحله‌ای بهبود یافته، با سایر سیستم‌های تشخیص نفوذ، پیاده‌سازی‌هایی انجام شده و نتایج حاصل برای ۵ گروه عادی، انکار سرویس، کاربر از راه دور، کاربر به ریشه و پویش‌گر، از نظر معیارهای دقت و نرخ تشخیص، به ترتیب در جدول‌های (۱۱) و (۱۲) و از نظر معیارهای صحت، بازخوانی و معیار F، به‌ترتیب در شکل‌های (۹)، (۱۰) و (۱۱) قابل مشاهده هستند.



شکل (۷): مقایسه میانگین معیار دقت (برای پنج گروه عادی، انکار سرویس، کاربر از راه دور، کاربر به ریشه و پویش‌گر) پس از ۱۰۰، ۲۰۰، ۵۰۰، ۱۰۰۰، ۱۰۰۰۰، ۲۰۰۰ و ۵۰۰۰۰ بار تکرار الگوریتم، به ازای توزیع‌های مختلف

Figure (7): Comparison of the mean Accuracy criterion (for five classes normal, DoS, R2L, U2R, and Probe) after 100, 200, 500, 1000, 10000, 20000, and 50000 repetitions of the algorithm, for different distributions.

Table (9): Comparing the clustering with classification

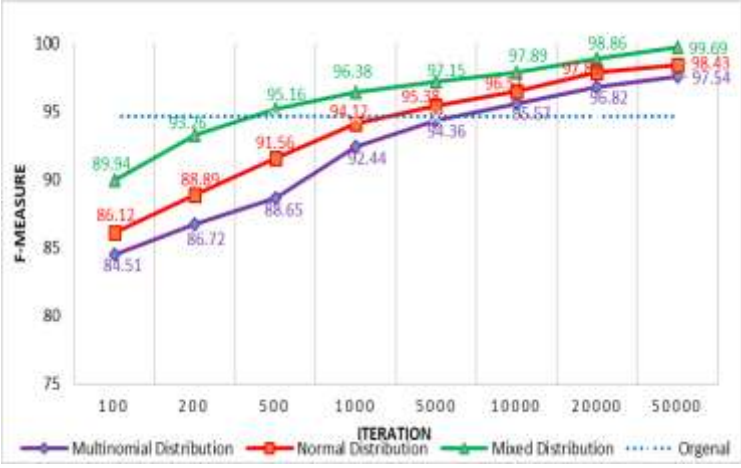
جدول (۹): مقایسه خوشه‌بندی با دسته‌بندی

الگوریتم	تعداد ویژگی‌ها	دقت (درصد)	مدت زمان آموزش (ثانیه)	مدت زمان آزمایش (ثانیه)
شبکه خودسازمان‌ده	۴۱	۹۸/۱۸	۱۳/۲۰	۰/۱۴
ماشین بردار پشتیبان	۴۱	۹۷/۸۲	۵۴/۱۸	۰/۲۵

Table (10): Comparing Two-Step clustering with individual clustering

جدول (۱۰): مقایسه خوشه‌بندی تجمیعی دومرحله‌ای با خوشه‌بندهای منفرد

روش خوشه‌بندی	دقت (درصد)	نرخ تشخیص (درصد)	نرخ اعلام نادرست (درصد)	بازخوانی (درصد)	صحت (درصد)	معیار F (درصد)
K میانگین	۷۷/۲۵	۸۲/۳۵	۵/۲۱	۸۳/۳۹	۷۶/۵۰	۷۹/۸۰
فازی C میانگین	۸۲/۱۳	۸۴/۶۷	۴/۲۴	۸۴/۸۸	۸۵/۱۲	۸۴/۹۹
Y میانگین	۸۷/۱۵	۸۶/۳۲	۳/۹۱	۸۹/۰۵	۸۹/۰۵	۸۹/۰۵
K واسط	۹۶/۳۸	۹۱/۲۸	۳/۲۲	۹۶/۴۴	۹۵/۱۷	۹۵/۸۰
ژنتیک	۹۷/۳۹	۹۵/۶۲	۴/۳۷	۹۷/۲۱	۹۸/۰۶	۹۷/۶۳
شبکه عصبی خودسازمان‌ده	۹۸/۱۸	۹۷/۰۹	۳/۱۳	۹۸/۱۳	۹۸/۴۲	۹۸/۲۷
روش پیشنهادی	۹۹/۸۷	۹۹/۷۳	۱/۰۵	۹۹/۴۴	۹۹/۶۷	۹۹/۵۵



شکل (۸): مقایسه میانگین معیارF (برای پنج گروه عادی، انکار سرویس، کاربر از راه دور، کاربر به ریشه و پویش‌گر) پس از ۱۰۰، ۲۰۰، ۵۰۰، ۱۰۰۰، ۱۰۰۰۰، ۲۰۰۰۰ و ۵۰۰۰۰ بار تکرار الگوریتم، به ازای توزیع‌های مختلف

Figure (8): Comparison of the mean F-Measure criterion (for five classes Normal, DoS, R2L, U2R, and Probe) after 100, 200, 500, 1000, 10000, 20000, and 50000 repetitions of the algorithm, for different distributions

جدول (۱۱): مقایسه معیار ارزیابی دقت روش پیشنهادی با سایر روش‌های تشخیص نفوذ

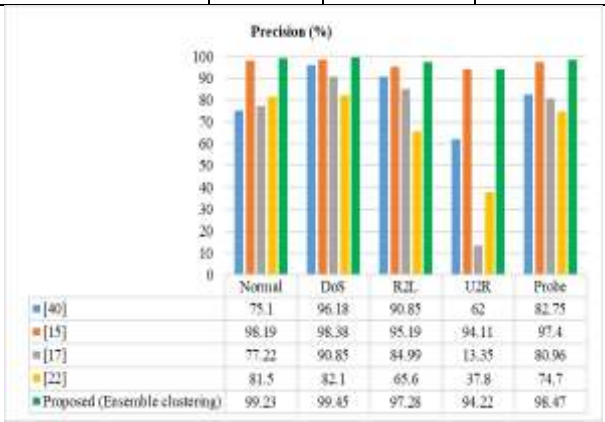
Table (11): Comparing the Accuracy of the proposed method with other intrusion detection methods

رویکردها	دقت برحسب درصد				
	عادی	انکار سرویس	کاربر از راه دور	کاربر به ریشه	پویش‌گر
CRF و الگوریتم oneR [15]	۹۸/۵۸	۹۸/۰۲	۹۶/۱۱	۹۲/۳۰	۹۶/۵۷
ABC و الگوریتم تقویت تطبیقی [۱۶]	۹۸/۷۱	۹۹/۸۶	۹۷/۹۰	۹۸/۸۶	۹۹/۱۸
شبکه عصبی چند کانولوشنی]۱۷[	۸۴/۶۲	۹۲/۶۹	۹۱/۳۲	۹۷/۹۷	۹۶/۰۶
گام کمند خطی نزدیک‌ترین همسایه]۱۸[	۹۷/۱۴	۹۷/۰۱	۹۳/۵۶	۹۵/۶۷	۹۸/۲۴
روش پیشنهادی (خوشه‌بندی تجمیعی)	۹۹/۸۵	۹۹/۸۶	۹۸/۲۴	۹۸/۸۷	۹۹/۳۳

Table (12): Comparing the Detection Rate of the proposed method with other intrusion detection methods

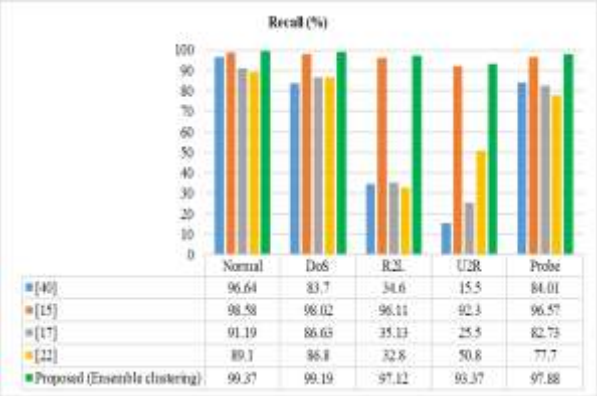
جدول (۱۲): مقایسه معیار ارزیابی نرخ تشخیص روش پیشنهادی با سایر روش‌های تشخیص نفوذ

رویکردها	نرخ تشخیص برحسب درصد				
	نرمال	انکار سرویس	کاربر از راه دور	کاربر به ریشه	پویش‌گر
یادگیری مهاجرت عمیق [۱۹]	۹۸/۶۸	۹۹/۹۱	۸۳/۳۳	۸۳/۳۳	۹۶/۶۳
ABC و الگوریتم تقویت تطبیقی [۱۶]	۹۹/۹۸	۹۹/۹۱	۹۹/۳۷	۹۹/۳۷	۹۹/۸۹
بِیز ساده و عامل قطعیت(CF-KNN) KNN [۲۰]	۹۴/۵۶	۸۴/۶۸	۶۷/۱۶	۶۷/۱۶	۷۹/۷۶
ماشین یادگیری افراطی چند هسته‌ای 21[(MK-ELM)]	۹۹/۱۲	۰۳.۹۸	۷۶/۱۵	۷۶/۱۵	۹۵/۷۴
روش پیشنهادی (خوشه‌بندی تجمیعی)	۹۹/۸۴	۹۹/۹۲	۹۸/۴۳	۹۸/۴۳	۹۹/۹۱



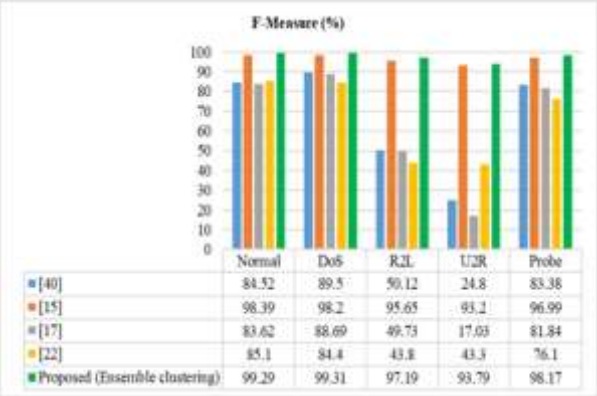
شکل (۹): نمودار مقایسه معیار ارزیابی صحت روش پیشنهادی با سایر روش‌های تشخیص نفوذ

Figure (9): The diagram for comparing the proposed method with other intrusion detection methods, in terms of Precision



شکل (۱۰): نمودار مقایسه معیار ارزیابی بازخوانی روش پیشنهادی با سایر روش‌ها تشخیص نفوذ

Figure (10): The diagram for comparing the proposed method with other intrusion detection methods, in terms of Recall



شکل (۱۱): نمودار مقایسه ارزیابی معیار F روش پیشنهادی با سایر روش‌های تشخیص نفوذ

Figure (11): The diagram for comparing the proposed method with other intrusion detection methods, in terms of F-Measure

۶- نتیجه گیری

توسعه تکنولوژی انتقال صدا از طریق پروتکل اینترنت و افزایش سوء استفاده از آسیب‌پذیری‌های آن باعث می‌شود امنیت و به ویژه، تشخیص نفوذ در این شبکه‌ها، به‌عنوان موضوعات مهم در نظر گرفته شوند. با توجه به این الزامات، تلاش‌های متعددی جهت طراحی سیستم‌های تشخیص نفوذ برای شبکه‌های VoIP متمرکز شده‌اند که بیشتر آنها مبتنی بر رویکردهای یادگیری ماشین هستند.

با توجه به اهمیت ویژگی‌های ترافیکی، برای جداسازی بسته‌های عادی و بسته‌های حاوی حمله‌های مختلف، شناسایی و انتخاب ویژگی‌های برتر برای تشخیص نفوذ در شبکه‌های VoIP، به‌عنوان اولین چالش مورد نظر قرار گرفته و الگوریتم شبیه‌سازی تبرید و در کنار آن، یک مدل طبقه‌بندی مبتنی بر MLP، به‌عنوان ارزیاب، برای حل این چالش پیشنهاد شده است. به‌عنوان راه‌حلی برای چالش دوم، یعنی توجه به روابط بین نمونه‌ها برای رهایی از نقاط پرت، یک مدل خوشه‌بند تجمیعی مبتنی بر نسخه بهبود یافته الگوریتم دومرحله‌ای در نظر گرفته شده است. ارزیابی رویکرد پیشنهادی برای انتخاب ویژگی و مقایسه آن با رویکردهای ارائه شده قبلی، از نظر ابزارهای بهینه‌سازی، بهبود متوسط ۲۳/۹۱ درصد برای زمان آموزش و ۱۸/۵ درصد برای زمان آزمون را نشان می‌دهد. همچنین استفاده از MLP به‌عنوان ارزیاب، باعث بهبود متوسط ۲۲/۵ درصد و ۳۳ درصد، به‌ترتیب در زمان آموزش و آزمون می‌شود. علاوه بر این، مقایسه رویکرد پیشنهادی، مبتنی بر خوشه‌بندی دومرحله‌ای بهبود یافته، میانگین بهبود دقت، نرخ تشخیص و معیارF را به‌ترتیب، معادل۳/۳۴ درصد، ۱۷/ ۱۴درصد و ۳۲/۸۷ درصد نشان می‌دهد.

سپاسگزاری

این مقاله مستخرج از رساله دوره دکتری در دانشگاه آزاد اسلامی واحد دزفول است. نویسندگان بر خود لازم می‌دانند مراتب تشکر صمیمانه خود را از همکاران حوزه پژوهشی دانشگاه آزاد اسلامی و داوران محترم که ما را در انجام و ارتقای کیفی این مقاله یاری نموده‌اند، اعلام نمایند.

References

مراجع

[1] S. Armoogum, N. Mohamudally, "An extended genetic algorithm-based prevention system against DoS/DDoS flood attacks in VoIP systems", Progress of the Springer/ACIE, pp. 301-312, Singapore, Apr. 2021 (doi: 10.1007/978-981-33-4299-6_25).

[2] V. Kumar, O.P. Roy, "Reliability and security analysis of VoIP communication systems", InRising Threats in Expert Applications and Solutions, Proceeding of the Springer/FICR-TEAS, pp. 687-693, Singapore, 2021 (doi: 10.1007/978-981-15-6014-9_84).

[3] Q. Wang, Q. Qian, "Malicious code classification based on opcode sequences and textCNN network", Journal of Information Security and Applications, vol. 67, Article Number: 103151, June 2022 (doi: 10.1016/j.jisa.2022.103151).

[4] A. Salman, M.S. Khan, S. Idrees, F. Akram, M. Junaid, AL. Malik, "File integrity checkers: functionality, attacks, and protection", Proceeding of the IEEE/ICoDT2, pp. 1-6, Rawalpindi, Pakistan, May 2022 (doi: 10.1109/ICoDT255437.2022.9787428).

[5] F. Salo, M. Injadat, A. Moubayed, A.B. Nassif, A. Essex, " Clustering enabled classification using ensemble feature selection for intrusion detection", Proceeding of the IEEE/ICNC, pp. 276-281, Honolulu, HI, USA, Feb. 2019 (doi: 10.1109/ICCNC.2019.8685636).

[6] K.K. Wankhade, K.C. Jondhale, "An ensemble clustering method for intrusion detection", International Journal of Intelligent Engineering Informatics, vol. 7, no. 2-3, pp. 112-140, April. 2019 (doi: 10.1504/IJIEI.2019.099-085).

[7] X. Wu, T. Ma, J. Cao, Y. Tian, A. Alabdulkarimm, "A comparative study of clustering ensemble algorithms", Computers and Electrical Engineering, vol. 68, pp. 603-615, May. 2018 (doi: 10.1016/j.compeleceng.2018.05.005).

[8] S. Khanmohammadi, N. Adibeig, S. Shanehbandy, "An improved overlapping k-means clustering method for medical applications", Expert Systems with Applications, vol. 67, pp. 12-18, Jan. 2017 (doi: 10.1016/j.eswa.2016.09.025).

[9] T. Chiu, D. Fang, G. Chen, Y. Wang, C. Jeris, "A robust and scalable clustering algorithm for mixed type attributes in large database environment", Proceeding of the seventh ACM SIGKDD, pp. 263-268, San Francisco California, Aug. 2001 (doi: 10.1145/502512.502549).

[10] J. Bacher, K. Wenzig, M. Vogler, "SPSS TwoStep Cluster- A first evaluation", Friedrich-Alexander University of Erlangen-Nuremberg, Chair of Sociology, vol. 2, pp. 1-23, Jan. 2004.

[11] V.R. Balasaraswathi, M. Sugumaran, Y. Hamid, "Feature selection techniques for intrusion detection using non-bio-inspired and bio-inspired optimization algorithms", Journal of Communications and Information Networks, vol. 2, no. 4, pp. 107-119, Dec. 2017 (doi: 10.1007/s41650-017-0033-7).

[12] M. Sharma, P. Kaur, "A comprehensive analysis of nature-inspired meta-heuristic techniques for feature selection problem", Archives of Computational Methods in Engineering, vol. 28, pp. 1103-1127, May 2021 (doi: 10.1007/s11831-020-09412-6).

[13] M. Abd Elaziz, A.H. Elsheikh, D. Oliva, L. Abualigah, S. Lu, A.A. Ewees, "Advanced metaheuristic techniques for mechanical design problems: Review", Archives of Computational Methods in Engineering, vol. 29, pp. 695–716, Jan. 2022 (doi: 10.1007/s11831-021-09589-4).

[14] P.J. Van Laarhoven, E.H. Aarts, "Simulated annealing", in simulated annealing: Theory and applications, Springer, Netherlands, vol. 37, pp. 7-15, 1987 (doi: 10.1007/978-94-015-7744-1_2).

[15] A. Mahendiran, R. Appusamy, "An intrusion detection system for network security situational awareness using conditional random fields", International Journal of Intelligent Engineering and Systems, vol. 11, no. 3, pp. 196-204, June 2018 (doi: 10.22266/ijies2018.0630.21).

[16] M. Mazini, B. Shirazi, I. Mahdavi, "Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms", Journal of King Saud University-Computer and Information Sciences, vol. 31, no. 4, pp. 541-553, Oct. 2019 (doi: 10.1016/j.jksuci.2018.03.011).

[17] Y. Li, Y. Xu, Z. Liu, H. Hou, Y. Zheng, Y. Xin, Y. Zhao, L. Cui , "Robust detection for network intrusion of industrial IoT based on multi-CNN fusion", Measurement, vol. 154, pp. 107450, Mar. 2020 (doi: 10.1016/j.-mea surement.2019.107450).

[18] X. Li, P. Yi, W. Wei, Y. Jiang, L. Tian, "LNNLS-KH: A feature selection method for network intrusion detection", Security and Communication Networks, pp. 1-22, Jan. 2021 (doi: 10.1155/2021/8830431).
[19] D. Li, L. Deng, M. Lee, H. Wang, "IoT data feature extraction and intrusion detection system for smart cities based on deep migration learning", International Journal of Information Management, vol. 49, pp. 533-545, Dec. 2019 (doi: 10.1016/j.ijinfomgt.2019.04.006).
[20] H.H. Pajouh, G. Dastghaibyfard, S. Hashemi, "Two-tier network anomaly detection model: A machine learning approach", Journal of Intelligent Information Systems, vol. 48, pp. 61-74, Feb. 2017 (doi: 10.1007/s1-0844-015-0388-x).
[21] W. Zhang, D. Han, K.C. Li, F.I. Massetto, "Wireless sensor network intrusion detection system based on MK-ELM", Soft Computing, pp. 12361-12374, Aug. 2020 (doi: 10.1007/s00500-020-04678-1).
[22] P. Bedi, N. Gupta, V. Jindal, "I-SiamIDS: An improved Siam-IDS for handling class imbalance in network-based intrusion detection systems", Applied Intelligence, vol. 51, pp. 1133-1151, Feb. 2021 (doi: 10.1007/s10-489-020-01886-y).
[23] A. Chellam, L. Ramanathan, S. Ramani, "Intrusion detection in computer networks using lazy learning algorithm", Procedia Computer Science, vol. 132, pp. 928-936, Jan. 2018 (doi: 10.1016/j.procs.2018.05.108).
[24] A.C. Enache, V. Sgârciu, "Anomaly intrusions detection based on support vector machines with an improved bat algorithm", Proceeding of the IEEE/CSCS, pp. 317-321, Bucharest, Romania, May. 2015 (doi: 10.110-9/CSCS.2015.12).
[25] D. Jianjian, T. Yang, Y. Feiyue, "A novel intrusion detection system based on IABRBFSVM for wireless sensor networks", Procedia Computer Science, vol. 131, pp. 1113-1121, Jan. 2018 (doi: 10.1016/j.procs.2-018.04.275).
[26] E. Kabir, J. Hu, H. Wang, G. Zhuo, "A novel statistical technique for intrusion detection systems", Future Generation Computer Systems, vol. 79, pp. 303-318, Feb. 2018 (doi: 10.1016/j.future.2017.01.029).
[27] V. Jyothsna, K.M. Prasad, "Anomaly-based intrusion detection system", Computer and Network Security, vol. 2, pp. 35-51, June. 2019.
[28] W. Wang, J. Liu, G. Pitsilis, X. Zhang, "Abstracting massive data for lightweight intrusion detection in computer networks", Information Sciences, vol. 433, pp. 417-430, April 2018 (doi: 10.1016/j.ins.2016.-10.023).
[29] A. Shenfield, D. Day, A. Ayesh, "Intelligent intrusion detection systems using artificial neural networks", IET Express, vol. 4, no. 2, pp. 95-99, June. 2018 (doi: 10.1016/j.icte.2018.04.003).
[30] A.L. Fred, A.k. Jain, "Combining multiple clusterings using evidence accumulation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 835-850, July 2005 (doi: 10.1109/TPAMI.2-005.113).
[31] D. Huang, J.H. Lai, C.D. Wang, "Robust ensemble clustering using probability trajectories", IEEE Trans. on Knowledge and Data Engineering, vol. 28, no. 5, pp. 1312-1326, Nov. 2015 (doi: 10.1109/TKDE.2015.250-3753).
[32] D. Huang, C.D. Wang, H. Peng, J. Lai, C.K. Kwoh, "Enhanced ensemble clustering via fast propagation of cluster-wise similarities", IEEE Trans. on Systems, Man, and Cybernetics: Systems, vol. 51, no. 1, pp. 508-520, Nov. 2018 (doi: 10.1109/TSMC.2018.2876202).
[33] C. Shao, S. Ding, "Link-based cluster ensemble method for improved meta-clustering algorithm", Proceeding of the Springer/IIP, pp. 14-25, Hangzhou, China, July 2020 (doi: 10.1007/978-3-030-46931-3_2).
[34] N. Iam-On, T. Boongoen, S. Garrett, "LCE: a link-based cluster ensemble method for improved gene expression data analysis", Bioinformatics, vol. 26, no. 12, pp. 1513-1519. June. 2010 (doi: 10.1093/bioinfo-rmatics/btq226).
[35] E. Jaw, X. Wang, "Feature selection and ensemble-based intrusion detection system: An efficient and comprehensive approach", Symmetry, vol. 13, no. 10, pp. 1764, Oct. 2021 (doi: 10.3390/sym13101764).
[36] N. Iam-On, T. Boongoen, S. Garrett, C. Price, "A link-based approach to the cluster ensemble problem", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 33, no. 12, pp. 2396-2409, May. 2011 (doi: 10.1109/TPAMI.2011.84).
[37] N. Iam-On, T. Boongoen, S. Garrett, "Refining pairwise similarity matrix for cluster ensemble problem with cluster relations", Proceeding of the Springer/DS, pp. 222-233, Berlin, Heidelberg, Oct. 2008 (doi: 10.1007-/978-3-540-88411-8_22).
[38] L. Dhanabal, S.P. Shantharajah, "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms", International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 6, pp. 446-452, June 2015 (doi: 10.17148/IJARCCE.2015.4696).
[39] T. Saranya, S. Sridevi, C. Deisy, T.D. Chung, M.A. Khan, " Performance analysis of machine learning algorithms in intrusion detection system: A review", Procedia Computer Science, vol. 171, pp. 1251-1260, Jan. 2020 (doi: 10.1016/j.procs.2020.04.133).
[40] C. Yin, Y. Zhu, S. Liu, J. Fei, H. Zhang, "Enhancing network intrusion detection classifiers using supervised

adversarial training", The Journal of Supercomputing, vol. 76, no. 9, pp. 6690-6719, Sept. 2020 (doi: 10.10-07/s11227-019-03092-1).

زیرنویس‌ها

1. Packet switching
2.Ttransmission control protocol/Internet protocol
3. Voice over IP
4. Gateways
5. Voicemail
6. Fax
7. SoftPhone
8. Denial of Service (DoS)
9. Distributed denial of service
10. Smurf/Smurfing attack
11. Synchronization flood
12. User datagram protocol flood
13. Internet control message protocol flood
14. Dynamic host configuration protocol
15. Flash-crowd attack
16. Pharming
17. Toll fraud
18. Public switched telephone network
19. Monitoring
20. Metaheuristic
21. Ensemble clustering systems
22. Twostep
23. Robust
24. Simulated Annealing algorithm
25. Multi-layer perceptron
26. Exclusive clustring
27. Overlapping clustring
28. Hierarchical clustring
29. K-Means
30. Greedy algorithms
31. Stepwise-optimal
32. Top-down
33. Bottom-up
34. Single-link
35. Complete-link
36. Average-link
37. Akaike information criterion
38. Bayesian information criterion
39. Heuristic
40. Random method
41. Conditional random field
42. Network security lab-knowledge discovery in databases
43. Anomaly network-based IDS (A-NIDS)
44. Artificial bee colony (ABC)
45. AdaBoost (Adaptive Boost)
46. Detection rate
47. False positive rate
48. Fusion model
49. Multi-convolutional neural network (multi-CNN)
50. Low efficiency
51. False positive rate
52. Krill swarm algorithm
53. Linear nearest neighbor lasso step (LNNLS-KH)
54. Intelligent city
55. Deep migration learning model
56. K-nearest neighbors (KNN)
57. Wireless sensor network (WSN)
58. F-Measure
59. Lazy learning
60. Bat algorightm
61. Artificial bee colony
62. Particle swarm optimization (PSO)
63. Least square support vector machine
64. Affinity propagation
65. Shell code
66. Random walk
67. Integration
68. Enhanced co-association (ECA)
69. Link-based meta-clustering algorithm (L-MCLA)
70. Weighted connected-triple (WCT)
71. Link-based cluster ensemble (LCE)
72. Hybrid feature selection
73. Rule-based engine
74. OneClass support vector machine
75. Density based spatial clustering of applications with noise
76. Expectation-maximization
77. SimRank based similarity
78. Nominal
79. Mean squared error (MSE)
80. Pre-cluster
81. Log-likelihood
82. Normal
83. Remote to local (R2L)
84. User to root (U2R)
85. Probe
86. True positive (TP)
87. True negative (TN)
88. False positive (FP)
89. False negative (FN)
90. Accuracy
91. Detection rate
92. Recall
93. Precision
94. False alarm rate
95. Support vector machines (SVM)
96. Naïve bayes
97. Normal probability distribution
98. Polynomial probability distribution
99. Mixture probability distribution