ارائه یک روش تشخیص نفوذ هوشمند مبتنی بر رفتار بر روی ماشین‌های مجازی

فاطمه میرجلیلی و جعفر رزم‌آرا*

گروه علوم کامپیوتر، دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه تبریز، تبریز، ایران

چکیده

امروزه، فناوری مجازی‌سازی به‌طور گسترده‌ای در جهت راه‌اندازی چندین سامانه مجازی بر روی یک سامانه فیزیکی در حال توسعه است که ابرهای محاسباتی نمونه بارز به‌کارگیری این فناوری هستند. سامانه‌های تشخیص نفوذ نقش کلیدی در محافظت از منابع ابر بر روی ماشین‌های مجازی دارند. با افزایش سرعت و پیچیدگی این ماشین‌ها، سامانه‌های تشخیص نفوذ نیز باید توانایی و دقت تشخیص خود را بالا ببرند تا توانایی شناسایی انواع مختلف حملات در زمان مناسب را کسب کنند. در این راستا، استفاده از رویکردهای مبتنی بر رفتار به‌دلیل مقیاس‌پذیری بالا در شبکه‌های بزرگ مورد توجه قرار گرفته‌اند. در این مقاله، یک روش جدید برای تشخیص نفوذ در شبکه مبتنی بر رفتار ارائه شده است. برای این منظور، روش پیشنهادی ابتدا داده‌های استخراج‌شده را از طریق مفهوم جریان داده به‌صورت گراف پراکندگی ترافیک مدل‌سازی کرده و سپس، با استفاده از یک الگوریتم بهبود یافته مبتنی بر مدل مارکوف خوشه‌بندی می‌کند. در ادامه، با تحلیل خوشه‌های تولیدشده بر اساس معیارهای آماری مدلی برای تشخیص نفوذ ارائه می‌دهد. کارایی روش پیشنهادی بر روی مجموعه‌داده DARPA 99 به‌عنوان یک مجموعه‌داده استاندارد و جامع برای ارزیابی سامانه‌های تشخیص نفوذ مورد آزمایش و ارزیابی قرار گرفت و با نتایج حاصل از هفت روش مطرح دیگر مقایسه شد. نتایج مقایسه نشان می‌دهد که روش پیشنهادی در مقایسه با روش‌های دیگر دارای قابلیت‌های بهتری بوده و می‌تواند حملات را با دقت بالایی تشخیص دهد.

واژگان کلیدی: تشخیص نفوذ مبتنی بر رفتار، گراف پراکندگی ترافیک، جریان داده، خوشه‌بندی گراف، خوشه‌بندی بهبودیافته مارکوف

An intelligent behavior-based intrusion detection method for virtual machines

Fatemeh Mirjalili & Jafar Razmara*

Department of Computer Science, Faculty of Mathematics, Statistics, and Computer Science, University of Tabriz, Tabriz, Iran

Abstract

In recent years, the speed and complexity of computer networks have grown significantly. At the same time, network-based anomalies and attacks have increased. Nowadays, intrusion detection and prevention is considered as a main strategy in satisfying the security of computer systems and communication networks, and the detection of these attacks with high accuracy and the least error is very important, especially in the field of network management. Today, virtualization technology is widely developing in order to set up multiple virtual systems on a physical system. Computational clouds are the most hallmark of this technology. Intrusion detection systems play a key role in protecting cloud resources on virtual machines. An intrusion detection system has the task of monitoring events within a computer system and the communication networks, and detects unauthorized and abnormal behaviors to deal with them. The proposed systems for intrusion detection mainly use data mining, machine learning and statistical analysis of data. Therefore, it is natural that in some cases they lead to the production of false alarms. Consequently, it is essential to improve the accuracy and high detection capability of these systems. Regarding the increasing speed and complexity of these machines, it is necessary to increase the ability and accuracy of intrusion detection systems for identifying different types of attacks at a right time. In this regard, the

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

● تاریخ ارسال مقاله: ۱۳۹۷/۱۲/۱۷ ● تاریخ پذیرش: ۱۳۹۸/۸/۱۹ ● تاریخ انتشار: ۱۴۰۰/۰۷/۱۷ ● نوع مطالعه: پژوهشی

use of behavior-based approaches has attracted more attention due to their high scalability in the large networks. The methods for intrusion detection that utilize network traffic graph clustering do not have the accuracy and appropriateness with the speed of data transfer in the current computer networks. Thus, the solutions can be improved by choosing an appropriate strategy for clustering. In this paper, a new behavior-based method for detecting intrusion in computer networks is presented. To this end, the network data was modeled through the flow of data as a traffic dispersion graph and then clustered using an improved Markov-based algorithm. Then, by analyzing a set of statistical criteria, the produced clusters, a penetration detection model was constructed. A set of modified statistical criteria was defined and utilized for analyzing the constructed clusters. The proposed model was examined and evaluated on the DARPA 99 dataset. In addition, the results of the proposed method were compared with seven other methods which work based on machine learning techniques. The results show that in the proposed method, the error detection rate is significantly reduced and the accuracy rate of the method is increased compared to seven other intrusion detection approaches. The reason for this performance improvement can be attributed to the good performance of Markov's improved clustering algorithm, which has produced more accurate results on flow-based data. Also, defining and applying appropriate criteria to determine the threshold limits is effective in obtaining accurate results. In addition, the results demonstrate that the proposed model has better capabilities than the methods which are not use graph clustering and can detect attacks with high accuracy.

Keywords: behavior-based intrusion detection, traffic dispersion graph, data flow, graph clustering, optimized Markov clustering

۱- مقدمه

در سال‌های اخیر، سرعت و پیچیدگی شبکه‌های رایانه‌ای رشد به‌سزایی داشته و به موازات آن، ناهنجاری‌ها و حملات مبتنی بر شبکه نیز افزایش پیدا کرده‌اند. امروزه، تشخیص و جلوگیری از نفوذ به‌عنوان یک راه‌کار اصلی در برآورده‌کردن امنیت سیستم‌های رایانه‌ای و شبکه‌های ارتباطی مطرح و شناسایی این حمله‌ها با دقت بالا و کم‌ترین خطا به‌خصوص در حوزه مدیریت شبکه از اهمیت بالایی برخوردار است. یک سامانه تشخیص نفوذ وظیفه دارد تا بر وقایع درون سیستم رایانه‌ای و شبکه‌های ارتباطی بین آنها نظارت کند و رفتارهای غیرمجاز و ناهنجار را برای مقابله با آنها تشخیص دهد. سامانه‌های ارائه‌شده برای تشخیص نفوذ به‌طور عمده از روش‌های داده‌کاوی، یادگیری ماشین و تحلیل آماری داده‌ها بهره می‌برند و طبیعی است که در مواردی منجر به تولید هشدارهای اشتباه می‌شوند [3]؛ بنابراین، بهبود دقت و توانایی تشخیص بالای این سامانه‌ها بسیار ضروری است.

تاکنون، تلاش‌های متعددی برای طراحی و ارائه سامانه‌های تشخیص نفوذ با قابلیت‌های بالا انجام گرفته است. روش‌های ارائه‌شده از جنبه‌های مختلفی قابل دسته‌بندی هستند، ازجمله دسته‌بندی از لحاظ نحوه عملکرد الگوریتم مورد استفاده که به سه دسته روش‌های مبتنی بر امضا، روش‌های مبتنی بر رفتار غیرعادی و روش‌های پروتکل غیرعادی تقسیم می‌شوند [4]. از آن‌جایی که روش‌های مبتنی بر امضا براساس الگوهای حملات شناخته‌شده عمل می‌کنند قادر به شناسایی حملات جدید و ناشناخته نیستند؛ ازاین‌رو، روش‌های تشخیص نفوذ مبتنی بر رفتار برای شناسایی حملات جدید بهتر عمل کرده و بیشتر مورد توجه قرار گرفته‌اند [5].

به‌دلیل افزایش روزافزون حجم بسته‌های داده‌ای انتقالی از طریق شبکه‌های ارتباطی، بررسی محتوای تک‌تک بسته‌ها جهت شناسایی نفوذ کارآیی خود را از دست داده است. از جمله رویکردهایی که برای رفع این مشکل ارائه شده‌اند، استفاده از روش‌های تشخیص نفوذ مبتنی بر جریان است. این روش‌ها به جای تحلیل محتوای هر بسته، ویژگی‌های مشترک مجموعه‌ای از بسته‌ها را تحلیل می‌کنند. در سال ۲۰۱۰، اسپروتو [6] رویکردی برای تشخیص نفوذ بر اساس جریان داده در شبکه با استفاده از سری‌های زمانی مطرح ارائه داد. راه‌کار دیگری برای تشخیص نفوذ مبتنی بر جریان بر اساس تحلیل حمله توسط هلمونس و همکاران [7] در سال ۲۰۱۲ ارائه شده است. در این مطالعات، برتری و کارایی تشخیص نفوذ مبتنی بر جریان داده در شبکه نسبت به روش‌های مبتنی بر محتوا نشان داده شده است. در این میان، استفاده از الگوریتم‌های مبتنی بر گراف در افزایش دقت تشخیص نفوذ در شبکه‌های با مقیاس بزرگ بسیار مؤثر است [8, 9]. درواقع، ماهیت این الگوریتم‌ها متناسب با نیازهای مهم مطرح در این مسائل از جمله ضرورت نظارت بر ترافیک شبکه است. به‌منظور نظارت، تحلیل و به‌دست‌آوردن نمایی از ترافیک شبکه، گراف پراکندگی ترافیک شبکه در سال ۲۰۰۷ توسط ایلیوفوتو و همکاران [8] پیشنهاد شد. ایلیوفوتو گراف پراکندگی ترافیک را به‌عنوان نمایش گرافیکی از تعامل‌های مختلف بین گروه‌هایی از گره‌ها تعریف کرد. در شبکه‌های مبتنی بر پروتکل اینترنت، یک گره متعلق به موجودیتی است که یک نشانی پروتکل اینترنت خاص دارد و یک یال

نشان‌دهنده ارسال بسته بین دو گره متفاوت است. در این راستا، کوک‌له و همکارانش در سال ۲۰۱۱ [10] رویکرد جدیدی بر اساس گراف پراکندگی ترافیک شبکه برای تشخیص نفوذ و ناهنجاری‌ها مطرح کرند. در تلاش دیگری، مناندهر و آنگ در سال ۲۰۱۴ [11] رویکرد جدیدی بر اساس تحلیل سرآیند بسته‌ها به جای تحلیل محتوای بسته و با توانایی شناسایی حملات بر روی پروتکل TCP را ارائه دادند. با وجود میزان دقت تشخیص بالای این رویکرد، تحلیل سرآیند بسته‌ها در شبکه‌های با سرعت بالا کارایی خود را از دست داده است. در راه‌کار ارائه‌شده توسط راستگار و همکارانش در سال ۲۰۱۵ [1] خوشه‌بندی گراف جریان ترافیک شبکه اساس کار قرار گرفت. در این راه‌کار، به‌ازای هر سری زمانی از جریان ترافیک شبکه گراف ساخته شده است و با استفاده از یک روش مبتنی بر الگوریتم ژنتیک گراف‌های هر سری زمانی خوشه‌بندی و سپس ارزیابی می‌شود. این راه‌کار پیشنهادی میزان تشخیص اشتباه کمتری داشته و قادر به شناسایی انواع مختلفی از حملات، نسبت به رویکردهای مبتنی بر بسته است؛ اما استفاده از الگوریتم ژنتیک جهت خوشه‌بندی مسائلی مناسب است که محدودیت زمانی نداشته باشد و به عبارتی، خوشه‌بندی ژنتیک هر چه زمان بیشتری در اختیار داشته باشد، بهتر عمل می‌کند. در سامانه‌های تشخیص نفوذ شناسایی حملات باید در کمترین زمان ممکن انجام شود و در شبکه‌های با مقیاس بالا با میزان انتقال داده‌های زیاد الگوریتم ژنتیک نمی‌تواند کارا باشد. درمجموع، روش‌های ارائه‌شده بالا، با وجود استفاده از داده‌های مبتنی بر جریان و رویکردهای مبتنی بر گراف میزان دقت تشخیص ناهنجاری‌ها افزایش یافته اما این روش‌ها تنها قادر به شناسایی انواع محدودی از حملات هستند.

روش‌های ارائه‌شده برای تشخیص نفوذ که از خوشه‌بندی گراف ترافیک شبکه بهره می‌برند، دقت و تناسب لازم با سرعت انتقال داده‌ها در شبکه‌های امروز را ندارند و بنابراین، می‌توان با انتخاب راه‌کار مناسب برای خوشه‌بندی دقت و کیفیت راه‌کار را بهبود داد. در این مقاله، یک راه‌کار مبتنی بر خوشه‌بندی مارکوف برای تحلیل گراف ترافیک شبکه ارائه شده است. در همین راستا، برای هر سری زمانی از جریان ترافیک شبکه گراف ساخته شده و سپس خوشه‌بندی می‌شود. خوشه‌بندی براساس مدل مارکوف سریع عمل کرده و دارای دقت بالایی است. با توجه به سرعت و دقت بالای خوشه‌بندی مارکوف، معیارهای آماری موجود در جهت تحلیل نتایج خوشه‌بندی نیاز به تغییراتی دارند که در این مقاله مورد بررسی قرار گرفته است. در بخش بعدی، روش پیشنهادی معرفی شده و در بخش نتایج، ارزیابی روش پیشنهادی در مقایسه با روش‌های مشابه بررسی می‌شود.

۲- مواد و روش‌ها

بررسی روش‌های ارائه‌شده برای تشخیص نفوذ نشان می‌دهد که افزایش دقت در تشخیص نفوذ با محدودیت در تشخیص انواع حملات و سرعت تشخیص مواجه است. پژوهش حاضر با هدف اصلی ارائه روشی مناسب و کاربردی برای افزایش دقت تشخیص و سرعت خوشه‌بندی و همچنین کاهش محدودیت در تشخیص انواع حمله و نرخ هشدارهای غلط در سامانه‌های تشخیص نفوذ انجام گرفته است. به همین منظور با استفاده از مفاهیم جریان، گراف پراکندگی ترافیک و خوشه‌بندی آن رویکرد جدیدی معرفی شده است. قبل از تشخیص نفوذ لازم است، پیش‌پردازش‌های لازم بر روی ترافیک خام شبکه انجام گیرد تا داده‌ها برای روش ارائه‌شده قابل استفاده شوند. در این بخش، ابتدا روش مورد استفاده برای پیش‌پردازش داده‌ها معرفی و سپس، روش پیشنهادی برای تشخیص نفوذ شرح داده شده است.

۲-۱- پیش‌پردازش داده‌ها

یکی از جنبه‌های مهم در ارائه یک سامانه تشخیص نفوذ، اثبات کارآیی آن بر روی مجموعه‌داده‌های مناسب است. دو مجموعه CIDD [12] و DARPA [13] برای ارزیابی سامانه‌های تشخیص نفوذ مناسب هستند. مجموعه CIDD شامل منابع مجازی تهیه‌شده از محاسبات ابری بوده و از دو بخش داخلی و خارجی تشکیل شده است که بخش داخلی آن متشکل از ماشین‌های مجازی با سیستم عامل‌های مختلف و بخش خارجی مربوط به سیستم‌های کاربران است. بخشی از این مجموعه از روی مجموعه‌داده DARPA تهیه شده، ولی به‌دلیل عدم دسترسی کامل به این مجموعه و با توجه به تشابه کامل ساختار این دو مجموعه، در این پژوهش، مجموعه‌داده‌ای DARPA 99 برای ارزیابی روش پیشنهادی مورد استفاده قرار گرفت. این مجموعه‌داده مبتنی بر بسته بوده و با توجه به نبود دسترسی به مجموعه‌داده‌های مبتنی بر

جریان، داده‌های مبتنی بر بسته بالا به داده‌های مبتنی بر جریان طبق روندی که در ادامه توضیح داده شده تبدیل شد. برای این منظور، ابتدا ترافیک خام شبکه جمع‌آوری و سپس، سرآیند هر یک از بسته‌ها در مجموعه‌داده‌ای ثبت شد. داده‌های موجود به بازه‌های زمانی منظم و یکسان تقسیم و برای هر بازه زمانی، گراف مربوط به آن بازه و از روی گراف، ماتریس مجاورت آن ساخته شد.

هر سطر در مجموعه‌داده، نشان‌دهنده ویژگی‌های یک بسته بوده و به‌ترتیب، شامل شماره بسته، نشانی مبداء، نشانی مقصد، نام پروتکل، زمان ارسال، شماره درگاه مبداء و شماره درگاه مقصد است. با استفاده از این ویژگی‌ها، سرآیند هر بسته ساخته شد. جهت تسهیل در انجام مراحل پیاده‌سازی، نشانی‌های مبداء و مقصد به اعداد صحیح نگاشت داده شد. در هر ثانیه، بسته‌هایی که دارای ویژگی‌های مشترکی هستند در یک جریان قرار گرفته و سپس، تعداد بسته‌ها برای هر جریان محاسبه شد. شکل (۱) نمونه‌ای از مجموعه‌داده مبتنی بر جریان تهیه‌شده را نشان می‌دهد. هر سطر مجموعه‌داده‌ای به‌ترتیب شامل هفت ستون نشانی مبداء، نشانی مقصد، تعداد بسته‌های جریان موردنظر، زمان عبور جریان در شبکه، شماره درگاه مبداء، شماره درگاه مقصد و نوع پروتکل است.

```
2,1,1,08:00:04,123,0,UDP
1,2,1,08:00:04,123,0,UDP
950,3,1,08:00:05,520,0,UDP
950,3,1,08:00:31,520,0,UDP
4,2,1,08:00:48,53,0,UDP
2,4,1,08:00:48,53,0,UDP
6,5,3,08:00:48,1024,0,TCP
5,6,2,08:00:48,21,0,TCP
6,5,3,08:00:49,1024,0,TCP
5,6,2,08:00:49,21,0,TCP
5,6,2,08:00:50,21,0,TCP
6,5,2,08:00:50,1024,0,TCP
6,5,4,08:00:51,1024,0,TCP
5,6,3,08:00:51,21,0,TCP
5,6,5,08:00:51,20,0,TCP
```

(شکل-۱): نمونه‌ای از مجموعه‌داده مبتنی بر جریان
(Figure-1): An example of the flow-based dataset

۲-۲- ساخت گراف

در این مرحله، مجموعه‌داده ورودی به سری‌های زمانی منظم $t_1, t_2, t_3, \ldots, t_n$ تقسیم شده و به‌ازای هر بازه زمانی یک گراف $G = (V, E)$ ساخته می‌شود. گراف G گراف پراکندگی ترافیک نامیده می‌شود که یک گراف جهت‌دار بوده و در آن، V مجموعه‌ای از گره‌ها متشکل از نشانی اینترنتی نمونه‌ها در جریان‌های ترافیک شبکه و E نماینده مجموعه‌جریان‌های بین این نشانی‌ها است که از طریق یال‌ها گره‌ها را به هم وصل می‌کند. برای یک جفت گره، در صورت وجود جریان بین آن‌ها یک یال در نظر گرفته می‌شود. همچنین، تعداد جریان‌ها بین دو گره به‌عنوان وزن یال در نظر گرفته می‌شود. در روش پیشنهادی، بهترین سری زمانی برای تشخیص ناهنجاری‌ها با توجه به تراکم داده‌ها و سرعت شبکه انتخاب شد. برای تشخیص ناهنجاری‌ها، هر گراف در سری زمانی به‌صورت مستقل مورد بررسی قرار گرفت.

۲-۳- ساخت ماتریس مجاورت گراف

برای هر یک از گراف‌های به‌دست‌آمده در سری‌های زمانی مختلف یک ماتریس مجاورت ساخته می‌شود. در ماتریس مجاورت هر سطر و ستون نشان‌دهنده گره‌های موجود در آن سری زمانی هستند و مقادیر عناصر ماتریس را وزن یال‌های موجود بین دو گره تشکیل می‌دهد. به‌طوری‌که اشاره شد گراف ترافیک شبکه جهت‌دار است، ولی با تبدیل داده‌ها به مجموعه‌داده مبتنی برجریان، ارتباطات بین گره‌ها به حالت دوطرفه با وزن‌های مشابه یا متفاوت تبدیل می‌شود. البته به تعداد محدود ارتباط یک‌طرفه بین گره‌ها وجود دارد که در اصل یک‌طرفه نیستند؛ بلکه ارتباط معکوس آنها در خارج از بازه سری زمانی مورد نظر قرار گرفته است. با توجه به تعداد محدود این حالت‌ها و به‌منظور اخذ نتایج مطلوب از انجام خوشه‌بندی، می‌توان در طول خوشه‌بندی از آنها چشم‌پوشی کرد. به این ترتیب، گراف ترافیک شبکه را می‌توان در طول فرآیند خوشه‌بندی به‌صورت بدون جهت در نظر گرفت. در این گراف، برای ارتباطات دوطرفه بین گره‌ها، بیشترین وزن یال‌های بین دو گره به‌عنوان وزن یال مشترک در ماتریس مجاورت در نظر گرفته می‌شود و ارتباطات یک‌طرفه بین دو گره در ماتریس مجاورت درج نمی‌شود. شکل (۲)، نمونه‌ای از تبدیل ماتریس مجاورت گراف جهت‌دار به گراف بدون جهت را نشان می‌دهد.

$$\begin{bmatrix} 0 & 4 & 1 & 0 & 0 \\ 3 & 0 & 0 & 0 & 2 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 4 & 1 & 0 & 0 \\ 4 & 0 & 0 & 0 & 2 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \end{bmatrix}$$

(شکل-۲): نمونه‌ای از تبدیل ماتریس مجاورت گراف جهت‌دار به گراف بدون جهت

(Figure-2): An example for converting the directed neighbor graph matrix to undirected graph matrix

۴-۲- خوشه‌بندی گراف در سری زمانی با الگوریتم بهبود یافته مارکوف

الگوریتم خوشه‌بندی مارکوف توسط دانگن [14] در سال ۲۰۰۰ ارائه شد. این الگوریتم، یکی از روش‌های خوشه‌بندی مبتنی بر گراف است که با استفاده از دو عملگر ماتریسی به یافتن خوشه‌های طبیعی موجود در یک گراف می‌پردازد. اساس کار الگوریتم، جستجو بر روی ماتریس تلاقی درون یک گراف و شبیه‌سازی جریان‌های موجود در آن است. با توجه به اینکه الگوریتم خوشه‌بندی مارکوف سریع عمل کرده و با انجام اصلاحات در مراحل آن می‌توان خوشه‌های دقیق‌تری را ایجاد کرد، گراف‌های ساخته‌شده از ترافیک جریان داده با استفاده از الگوریتم خوشه‌بندی بهبودیافته مارکوف [2] و با اعمال تغییراتی در مراحل آن خوشه‌بندی شد. مراحل اجرای الگوریتم خوشه‌بندی به‌صورت زیر است:

۱) ورودی‌های الگوریتم: ماتریس مجاورت هر گراف به‌عنوان ورودی به الگوریتم داده می‌شود. مقدار ورودی پارامتر تورم برابر دو بوده و با توجه به استفاده از الگوریتم بهبودیافته مارکوف، پارامتر بسط به‌صورت حاصل‌ضرب ماتریس مجاورت اولیه در ماتریس میانی است. بیشترین مقدار درایه ماتریس مجاورت جهت محاسبه میزان پارامتر بسط، شناسایی شده و در متغیر Max_v درج می‌شود، تا در مرحله بسط مورد استفاده قرار گیرد.

۲) نرمال‌سازی کلی ماتریس: وزن یال‌های گراف نرمال‌سازی می‌شود. به این صورت که وزن هر درایه در ماتریس مجاورت بر مجموع وزن کل درایه‌ها در آن ماتریس تقسیم می‌شود؛ به این ترتیب، مقادیر تمامی درایه‌های ماتریس با مقداری بین بازه صفر و یک جایگزین خواهد شد که موجب ایجاد تعادل بین وزن درایه‌ها می‌شود.

$$M_{ij} = \frac{\text{وزن هر درایه}}{\text{مجموع کل وزن درایه‌های ماتریس}} \qquad (۱)$$

۳) نرمال‌سازی ستونی ماتریس: وزن هر درایه در ستون ماتریس مجاورت بر مجموع وزن کل درایه‌های آن ستون تقسیم می‌شود تا جمع مقادیر هر ستون برابر یک شود.

$$M_{ij} = \frac{\text{وزن هر درایه از ستون } x}{\text{مجموع کل وزن درایه‌های ستون } x} \qquad (۲)$$

۴) عمل گسترش روی ماتریس: با استفاده از الگوریتم بهبودیافته مارکوف، پارامتر گسترش به‌صورت حاصل‌ضرب ماتریس مجاورت اولیه در ماتریس میانی: $M * M_G$ محاسبه می‌شود. گراف‌های مجموعه‌داده، شامل انواع مختلف ماتریس با مقادیر درایه‌های متفاوت هستند. درحالتی که گراف پیچیده باشد، وزن یال‌ها نیز زیاد خواهد بود؛ بنابراین جهت اخذ نتیجه بهتر، پارامتر گسترش نسبت به گراف‌های متفاوت متغیر و پویا بوده و به‌صورت زیر محاسبه می‌شود:

$$Power = \begin{cases} 2 & Max_v < 10 \\ \frac{Max_v}{10} + Mod_{10} + 1 & Max_v > 10 \end{cases} \qquad (۳)$$

متغیر Power میزان پارامتر بسط و Max_v همان بیشترین مقدار درایه در ماتریس اولیه مجاورت است. متغیر Mod_{10} وجود یا نبود وجود باقی‌مانده Max_v بر ده را نشان می‌دهد، اگر باقی‌مانده برابر صفر باشد، مقدار Mod_{10} برابر صفر بوده و در غیر این صورت برابر یک خواهد بود.

پس از به‌دست‌آوردن پارامترهای لازم فرمول گسترش ماتریس به حالت زیر تغییر می‌کند:

$$Matrix = M^2 * \ M_G{}^{Power} \qquad (۴)$$

که در آن، M ماتریس میانی و M_G ماتریس اولیه مجاورت یا تلاقی است.

۵) عمل تورم روی ماتریس: بر اساس رابطه زیر محاسبه می‌شود:

$$(\Gamma_r \mathrm{M})_{pq} = (M_{pq})^r / \sum_{i=1}^{k} (M_{iq})^r \qquad (۵)$$

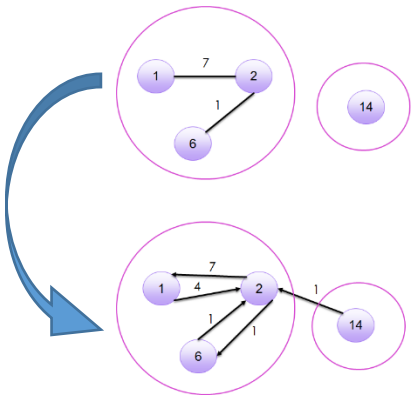
که در آن، متغیر r پارامتر تورم بوده و دارای مقدار دو است. نحوه محاسبه به این صورت است که تمام درایه‌های هر ستون ماتریس به توان دو رسیده، سپس مقدار هر درایه بر مجموع کل هر ستون تقسیم و در همان درایه درج می‌شود.

۶) مراحل ۴ و ۵ تا زمانی که الگوریتم به هم‌گرایی برسد تکرار می‌شود. الگوریتم زمانی به هم‌گرایی می‌رسد که ماتریس بدون تغییر باقی بماند. برای جلوگیری از احتمال گیرافتادن الگوریتم در حلقه تکرار، شرط تعداد تکرار حلقه یا همان تعداد نسل نیز با توجه به اندازه ماتریس تعیین می‌شود. در این مسأله بیشینه تعداد نسل برابر ۲۵ در نظر گرفته شده است.

۷) عمل هرس روی درایه‌های ماتریس: پس از خروج از حلقه، جمع مقادیر هر ستون ماتریس برابر یک بوده و در بیش‌تر موارد مقادیر هر درایه شامل یکی از اعداد صفر، یک و یا نیم است؛ اما در برخی حالات که الگوریتم با شرط تعداد نسل به هم‌گرایی رسیده مقادیر برخی درایه‌ها به‌صورت اعشاری و نزدیک به مقادیر صفر، یک و یا نیم خواهد بود که در این صورت اعداد گرد می‌شوند.

۸) تفسیر خوشه‌ها از روی ماتریس نهایی: برای تفسیر خوشه‌ها، ابتدا سطرهای صفر و سطرهای تکراری ماتریس حذف می‌شوند و سطرهای باقی‌مانده هرکدام نشان‌گر یک خوشه هستند. هر درایه از یک سطر که مقدارش بزرگتر از صفر باشد، یعنی آن گره به خوشه همان سطر متعلق است.

پس از خوشه‌بندی، یال‌ها به‌صورت جهت‌دار و وزن یال‌ها از روی گراف اولیه به‌روز می‌شوند. به عبارتی ارتباطات یک‌طرفه که قبل از خوشه‌بندی از آن‌ها چشم‌پوشی شده بود، دوباره در نظر گرفته خواهند شد و برای ارتباطات دوطرفه ناهماهنگ که قبل خوشه‌بندی هماهنگ شده بودند، دوباره هر یال، وزن خودش را خواهد داشت. شکل (۳)، نمونه‌ای از خوشه‌بندی بر روی مجموعه‌داده DARPA 99 است.



(شکل-۳): نمونه‌ای از خوشه‌بندی گراف
(Figure-3): A sample of the graph clustering

۵-۲- الگوریتم تشخیص نفوذ

پس از خوشه‌بندی گراف در هر سری زمانی، ویژگی‌های زیر برای آن محاسبه می‌شوند:

- زمان: این ویژگی با توجه به سری زمانی $t_1, t_2, t_3, \ldots, t_n$ با اعداد 1 تا n نشان داده می‌شود و مشخص می‌کند که ویژگی‌های ایجادشده مربوط به کدام سری زمانی هستند.
- شماره خوشه: خوشه‌های به‌وجودآمده با شماره نشان داده می‌شوند. شماره خوشه نشان می‌دهد که ویژگی‌ها مربوط به کدام خوشه در یک زمان مشخص هستند.
- تعداد گره‌های خوشه: این ویژگی تعداد گره‌های (پروتکل‌های اینترنت) موجود در یک خوشه را نشان می‌دهد و از یک بازه زمانی به بازه دیگر باهم متفاوت هستند.
- تعداد جریان‌های داخلی خوشه: تعداد جریان‌های موجود بین گره‌های یک خوشه را نشان می‌دهند.
- تعداد جریان‌های خارجی خوشه: تعداد جریان‌های موجود بین گره‌های دو خوشه متفاوت را مشخص می‌کند.

به این ترتیب مجموعه‌داده مبتنی بر خوشه‌بندی گراف با ویژگی‌های یادشده به‌دست می‌آید. خوشه با بیش‌ترین یال داخلی نشان‌دهنده اتصالات زیاد بین پروتکل‌های اینترنت است و این می‌تواند حاکی از حمله‌هایی از نوع عدم پذیرش سرویس، دسترسی محلی و یا کاربر ریشه‌ای باشد؛ همچنین، خوشه با بیش‌ترین یال خارجی نشان‌دهنده ارتباط یک پروتکل اینترنتی با چندین پروتکل اینترنتی دیگر است که این موضوع حمله‌هایی از نوع پویش را نشان می‌دهد.

در یک سامانه تشخیص نفوذ مبتنی بر رفتار در شبکه برای تشخیص رفتارهای ناهنجار از رفتارهای عادی، نیاز به معیارهای آماری وجود دارد. با استفاده از مفاهیم شرح‌داده‌شدهٔ بالا و همچنین ویژگی‌هایی که در مجموعه‌داده مبتنی بر خوشه‌بندی گراف وجود دارد، در ادامه با استفاده از معیارهای آماری ایستا که در [1] ارائه شده است، یک معیار جدید در جهت بهبود دقت تشخیص و همچنین کاهش نرخ مثبت کاذب پیشنهاد می‌شود.

- معیار مبتنی بر جریان داخلی خوشه: این معیار با استفاده از رابطه زیر محاسبه می‌شود:

$$AVGF_i = \frac{maxflow_i}{N_i} \qquad (۶)$$

که در آن $maxflow_i$ نشان‌دهنده تعداد جریان‌های داخلی خوشه‌ای است که بیش‌ترین تعداد جریان داخلی را دارد و N_i نشان‌دهنده تعداد گره در همان خوشه در سری زمانی i ام است.

- معیار مبتنی بر جریان خارجی خوشه: با تحلیل رفتار برخی از حمله‌ها مانند حمله‌های پویش، که در این نوع حمله‌ها جریان‌های خارجی بین بیش‌تر خوشه‌ها وجود دارد و مجموع مقادیر متوسط توزیع وزن خارجی سایر

خوشه‌ها بیشتر از بیشترین مقدار آن باشد. این معیار با استفاده از رابطه زیر محاسبه می‌شود:

$$AVGFE_i = \frac{maxF_i}{\sum_{j=1}^{k} F_j - maxF_i} \quad (۷)$$

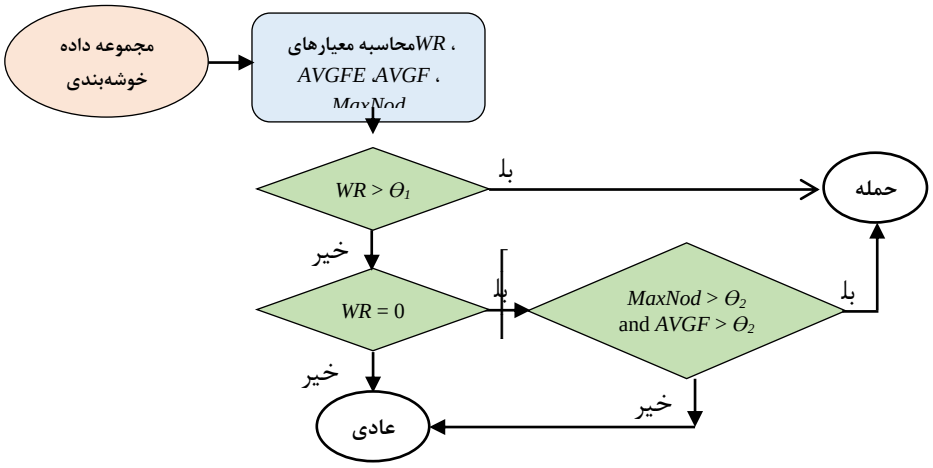
به‌طوری‌که $maxF_i$ بیشترین مقدار متوسط توزیع وزن خارجی در زمان i ام و F_j متوسط توزیع وزن خارجی در خوشه jام را نشان می‌دهد.

- معیار نسبت وزن‌ها: این معیار از نسبت معیار مبتنی بر جریان‌های داخلی خوشه به معیار مبتنی بر جریان‌های خارجی خوشه به‌دست می‌آید. اگر مقدار این معیار از یک حد آستانه مشخص بیشتر شود، یک ناهنجاری رخ می‌دهد. معیار نسبت وزن‌ها از رابطه زیر محاسبه می‌شود:

$$WR_i = \frac{AVGF_i}{AVGFE_i} \quad (۸)$$

- ترکیب معیار جریان داخلی خوشه با بیشترین جریان داخلی خوشه: برخی خوشه‌ها مستقل از سایر خوشه‌ها بوده و جریان خروجی به سایر خوشه‌ها ندارند. زمانی که حمله‌ای درون خوشه رخ می‌دهد، اگر جریان خارجی وجود نداشته باشد، میزان معیار نسبت وزن‌ها برابر صفر شده و از حد آستانه کمتر خواهد شد؛ درنتیجه حمله شناسایی نمی‌شود. برای حل این مشکل معیاری ترکیبی پیشنهاد می‌شود. برای این منظور، ابتدا معیارهای بالا محاسبه می‌شوند و اگر معیار نسبت وزن‌ها برابر صفر باشد، در داخل خوشه با بیشترین جریان داخلی، گرهی که نسبت به سایرین بیشترین ارتباط را دارد شناسایی و تعداد جریان‌های آن محاسبه می‌شود و در متغیر $MaxNod$ قرار می‌گیرد. درصورتی‌که مقدار متغیرهای $MaxNod$ و $AVGF_i$ از یک حدآستانه بیشتر شود، یک ناهنجاری رخ می‌دهد.

شکل (۴)، نمایی از مدل پیشنهادی جهت تشخیص نفوذ را نشان می‌دهد. در این مدل، ابتدا ترافیک خام شبکه مبتنی بر بسته از سرویس‌دهنده‌های درون شبکه گرفته شده و در مخزنی پردازش و نگه‌داری می‌شود. پردازش شامل دسته‌بندی بسته‌های دارای ویژگی یکسان در بازه‌های زمانی یک ثانیه است که هر کدام نشان‌گر یک جریان از داده هستند؛ سپس این مجموعه بر اساس سری زمانی مناسب به بازه‌هایی تقسیم شده، گراف‌های هر مجموعه ساخته شده و خوشه‌بندی می‌شوند و مجموعه‌داده مبتنی بر خوشه‌بندی را تشکیل می‌دهند. پس از محاسبه معیارها و ویژگی‌های مورد نیاز و حد آستانه‌های مناسب عمل تشخیص نفوذ انجام می‌شود.



(شکل-۴): روند تشخیص ناهنجاری‌ها با استفاده از روش پیشنهادی
(Figure-4): the procedure for Anomaly detection using the proposed method

۳- نتایج

به‌طوری که اشاره شد در این پژوهش، مجموعه‌دادهٔ DARPA 99 برای ارزیابی و اثبات کارآیی روش پیشنهادی مورد استفاده قرار گرفت. برای این منظور، ابتدا پارامترهای مدل پیشنهادی بر اساس داده‌های انتخابی از مجموعه بالا در طی مراحلی تنظیم شد؛ سپس، کارایی مدل در مقایسه با دو راه‌کار مطرح مورد مقایسه و ارزیابی قرار گرفت. در این بخش، نتایج حاصل شرح داده شده است.

۳-۱- تعیین حدآستانه نخست

مقدار حد آستانه اول (Θ_1) به‌صورت ایستا و بر مبنای مقایسه نتایج حاصل از نقاط عادی و نقاط ناهنجار تعیین شد. برای این منظور، داده‌های دو روز از هفته چهارم مجموعه‌داده DARPA 99 برای هر سه پنجره زمانی مورد

تحلیل قرار گرفته و بر اساس آن برای هر پنجره زمانی سه حد آستانه مشخص شد، سپس نتایج به‌دست‌آمده روی کل داده‌های هفته آزمایش و نتایج مورد بررسی و ارزیابی قرار گرفت.

۳-۲- انتخاب پنجره زمانی مناسب

برای انتخاب پنجره زمانی مناسب، مجموعه‌داده مبتنی بر جریان به سه بازه زمانی مساوی تقسیم شد. هر سری زمانی از بازه‌های $t_1, t_2, t_3, \ldots, t_n$ تشکیل شده به‌طوری‌که t_i نشان‌دهنده ۳۰، ۴۰ یا ۵۰ ثانیه i ام است؛ سپس، هر یک از بازه‌ها به‌ترتیب در مدل نهایی مورد آزمایش قرار گرفت. همچنین، برای انتخاب بهترین حالت، سه حد آستانه مختلف برای پارامتر *WR* در نظر گرفته شد. جدول (۱) نتایج حاصل به‌ازای سه بازه زمانی بالا را نشان می‌دهد. مقایسه نتایج سری‌های زمانی ۳۰، ۴۰ و ۵۰ ثانیه‌ای برای حد آستانه نخست نشان می‌دهد که معیارها در سری زمانی سی ثانیه و حد آستانه بزرگتر از سی بهترین نتایج را تولید کرده‌اند؛ بنابراین، حد آستانه دوم با درنظرگرفتن حد آستانه نخست بزرگ‌تر از سی در سری زمانی سی ثانیه محاسبه شد.

۳-۳- تعیین حدآستانه دوم

درصورتی‌که مقدار پارامتر *WR* کوچکتر از حد آستانه نخست باشد، برابربودن مقدار آن با صفر بررسی می‌شود. درصورتی‌که مقدار *WR* برابر صفر باشد، مقدار معیارهای *AVGF* و *MaxNod* بررسی می‌شود. اگر مقادیر آن‌ها از یک حد آستانه (Θ_2) بیشتر باشد، ناهنجاری رخ داده است. جدول (۲) نتایج حاصل از آزمایش‌های انجام گرفته با درنظرگرفتن سه مقدار مختلف برای حد آستانه دوم را نشان می‌دهد. مقایسه نتایج حاصل در این جدول برای حد آستانه دوم نشان می‌دهد که مقدار بیست بهترین نتایج را تولید می‌کند؛ پس از تعیین حد آستانه نخست و دوم، آزمایش بر روی کل داده‌های هفته چهارم انجام گرفت.

۳-۴- مقایسه و ارزیابی نتایج عملی

میزان کارایی راه‌کار پیشنهادی با مقایسه نتایج حاصل از آن با هفت روش مطرح دیگر مورد ارزیابی و بررسی قرار گرفت. داده‌های مورد استفاده برای آزمون روش‌ها مجموعه‌دادهٔ DARPA 99 هستند. این داده‌ها متشکل از یک مجموعه‌داده استاندارد است که به‌صورت عمومی قابل دسترس بوده و با آزمایش روش مورد نظر بر روی این مجموعه می‌توان قدرت و دقت یک روش را ارزیابی کرد.

به‌منظور ارزیابی و مقایسه روش پیشنهادی با سایر روش‌های مورد نظر، سه معیار دقت[1]، حساسیت[2] و ویژگی[3] مورد استفاده قرار گرفت. برای محاسبه این سه معیار، مقادیر نرخ مثبت (TP) صحیح و نرخ مثبت کاذب (FP) که به‌ترتیب، تعداد حملات صحیح و اشتباه تشخیص‌داده‌شده را نشان می‌دهد و نرخ منفی صحیح (TN) و نرخ منفی کاذب (FN) که به‌ترتیب، تعداد ارتباطات عادی صحیح و اشتباه تشخیص‌داده‌شده را شامل می‌شود، باید محاسبه شود. مقدار سه معیار بالا با استفاده از روابط زیر قابل محاسبه است:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (۹)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (۱۰)$$

$$Specificity = \frac{TN}{TN+FP} \quad (۱۱)$$

راه‌کارهای انتخابی برای مقایسه متشکل از تعداد پنج روش مبتنی بر روش‌های یادگیری ماشین شامل رگرسیون لجستیک[4]، ماشین بردار پشتیبان[5]، نیو بیز[6]، جنگل تصادفی[7] و درخت تصمیم گرادیان افزایشی[8] است که هر یک جزء روش‌های شناخته‌شده و کلاسیک در کاربردهای یادگیری ماشین هستند. نتایج درج‌شده در جدول (۳) حاصل مطالعه انجام‌گرفته توسط گوپتا و همکاران [15] است.

روش دیگر مورد مطالعه، روش ارائه‌شده توسط راستگار و همکارانش [1] است که یک الگوریتم مبتنی بر جریان با استفاده از معیارهای آماری ایستا است. در این روش، داده‌های مبتنی بر بسته به جریان تبدیل شده و سپس، خوشه‌بندی مبتنی بر الگوریتم ژنتیک بر روی مجموعه‌داده مبتنی بر جریان انجام می‌گیرد. با توجه به این موضوع، الگوریتم نیاز به زمان کافی جهت تولید نتایج خوب و بهینه دارد و بنابراین، در شبکه‌های با سرعت بالا که هر لحظه امکان وقوع حمله و ناهنجاری وجود دارد، ضعیف عمل می‌کند. در مقابل، الگوریتم خوشه‌بندی مارکوف سریع بوده و با توجه به تغییراتی که در الگوریتم

[1] Accuracy
[2] Sensitivity
[3] Specificity
[4] Logistic regression
[5] Support vector machine
[6] Naïve Bayes
[7] Random forest
[8] Gradient boosted decision tree

اولیه داده شده، نتایج دقیق‌تری را نسبت به زمانی که در اختیار دارد ارائه می‌دهد؛ به‌طوری‌که در بخش قبل شرح داده شد، در این پژوهش، معیارهای آماری استفاده‌شده در [1] تکمیل و بهبود داده شد. در جدول (۳)، نتایج درج‌شده برای روش مبتنی بر الگوریتم ژنتیک از مقاله [1] برگرفته شده است.

همچنین، کارایی روش پیشنهادی با روش تحلیل سرآیند بسته‌ها که در سال ۲۰۱۴ توسط مناندهر و آنگ [11] ارائه شد، مورد مقایسه و ارزیابی قرار گرفت. این روش توانایی شناسایی حملات مربوط به پروتکل TCP را دارد. نتایج ارائه‌شده در این روش به دو صورت نرخ هشدار کاذب پایین و نرخ هشدار کاذب بالا ارائه شده است. زمانی که میزان هشدار کاذب زیاد باشد، به‌علت پایین‌بودن حد آستانه‌ها، میزان نرخ مثبت صحیح افزایش پیدا می‌کند و در هشدار کاذب پایین به‌دلیل درنظرگرفتن حد آستانه بالا، میزان نرخ مثبت صحیح نیز کاهش پیدا می‌کند. نتایج درج‌شده در جدول (۳) برای روش تحلیل سرآیند بسته‌ها از مقاله [11] برگرفته شده است.

جدول (۳) ارزیابی نتایج حاصل از به‌کارگیری روش پیشنهادی و هفت روش تشخیص نفوذ را نشان می‌دهد. معیار دقت که بر اساس رابطه (۹) تعداد تشخیص‌های درست ارتباطات عادی و نفوذ را از بین تعداد کل موارد مورد بررسی نشان می‌دهد، مهم‌ترین معیار مطرح برای ارزیابی عملکرد هر یک از راه‌کارهای انتخابی است. در بین روش‌های مورد مطالعه روش پیشنهادی بهترین دقت تشخیص (۹۸/۶۱) را ارائه کرده و روش مبتنی بر الگوریتم ژنتیک (۹۶/۳۳) در رده دوم قرار دارد. روش پیشنهادی با انجام تغییراتی در مراحل الگوریتم خوشه‌بندی بهبودیافته مارکوف خوشه‌های دقیق‌تری را ایجاد کرده و با اعمال اصلاحات بر روی معیارهای آماری مورد استفاده در روش مبتنی بر الگوریتم ژنتیک موفق به کسب دقت بالاتری نسبت به سایر روش‌های مورد مطالعه شده است. همچنین، با توجه به کارایی بالاتر الگوریتم مارکوف نسبت به الگوریتم ژنتیک در خوشه‌بندی سریع گراف جریان داده، ارجحیت روش پیشنهادی ثابت می‌شود. همچنین، دو معیار حساسیت و ویژگی طبق رابطه (۱۰ و ۱۱) به‌ترتیب معرف نرخ تشخیص صحیح حملات از بین کل تعداد حملات و نرخ تشخیص صحیح ارتباطات عادی از بین کل تعداد این نوع ارتباطات انجام گرفته است. بر این اساس، روش پیشنهادی (۹۷/۷۲) در رده دوم بهترین نرخ حساسیت پس از روش نیو بیز (۹۹/۴۱) قرار گرفته اما نرخ ویژگی خیلی کم روش نیو بیز (۵۸/۵۶) باعث کسب دقت تشخیص پایین این روش شده است. از سوی دیگر، نرخ ویژگی کسب‌شده به‌وسیلهٔ روش پیشنهادی (۹۹/۵۳) با اندک تفاوتی نسبت به روش درخت تصمیم گرادیان افزایشی (۹۹/۷۹) در رده دوم و بالاتر از دو روش رگرسیون لجستیک (۹۸/۳۱) و تحلیل سرآیند بسته‌ها (۹۸/۲۵) قرار گرفته است. درمجموع، روش پیشنهادی با توجه به کسب نرخ بالای حساسیت و ویژگی موفق به اخذ بالاترین دقت تشخیص در بین روش‌های مورد مطالعه شده است.

(جدول-۱): دقت نتایج حاصل از به‌کارگیری سه پنجره زمانی و سه مقدار مختلف برای حد آستانه نخست
(Table-1): The obtained accuracy using three different timing window and three different values for the first threshold

پنجره زمانی	حد آستانه *WR*	نرخ مثبت کاذب	نرخ مثبت صحیح	نرخ منفی کاذب	نرخ منفی صحیح	دقت تشخیص
۳۰ ثانیه	۳۰ >	۰/۰۰۰۶	۰/۷۴	۰/۲۵	۰/۹۹	۰/۹۴
	۵۰ >	۰/۰۰۰۳	۰/۴۷	۰/۵۲	۰/۹۹	۰/۸۸
	۶۰ >	۰/۰	۰/۴۳	۰/۵۶	۰/۹۹	۰/۸۷
۵۰ ثانیه	۳۰ >	۰/۰۰۰۷	۰/۷۳	۰/۲۶	۰/۹۹	۰/۹۴
	۵۰ >	۰/۰۰۰۳	۰/۱۸	۰/۸۱	۰/۹۹	۰/۸۲
	۶۰ >	۰/۰۰۰۳	۰/۱۲	۰/۹۰	۰/۹۹	۰/۸۰
۶۰ ثانیه	۳۰ >	۰/۰۰۰۹	۰/۶۴	۰/۳۵	۰/۹۹	۰/۹۲
	۵۰ >	۰/۰۰۰۷	۰/۵۱	۰/۴۹	۰/۹۹	۰/۸۹
	۶۰ >	۰/۰	۰/۴۹	۰/۵۰	۰/۱	۰/۹۶

(جدول-۲): دقت نتایج حاصل از به‌کارگیری پنجره زمانی سی ثانیه و حد آستانه نخست بزرگ‌تر از سی
(Table-2): The obtained accuracy using timing window (=30 s) and the first threshold (>30)

حد آستانه دوم	نرخ مثبت کاذب	نرخ مثبت صحیح	نرخ منفی کاذب	نرخ منفی صحیح	دقت تشخیص
۲۰ >	۰/۰۰۵	۰/۹۷۷	۰/۰۲۳	۰/۹۹۵	۰/۹۸۶
۲۵ >	۰/۰۰۲	۰/۹۴۹	۰/۰۵	۰/۹۹	۰/۹۷۰
۳۵ >	۰/۰۰۱	۰/۸۰	۰/۱۹۹	۰/۹۹	۰/۹۵۶

(جدول-۳): مقایسه نتایج حاصل از روش پیشنهادی و سایر روش‌های تشخیص نفوذ
(Table-3): Comparing the results of the proposed method and other intrusion detection method

	حساسیت (Sensitivity)	ویژگی (Specificity)	دقت (Accuracy)
رگرسیون لجستیک	۹۰/۰۲	۹۸/۳۱	۹۱/۶۴
ماشین بردار پشتیبان	۹۲/۱۷	۹۱/۹۶	۹۲/۱۳
نیو بیز	۹۹/۴۱	۵۸/۵۶	۹۱/۴۵
جنگل تصادفی	۹۰/۸۵	۹۷/۴۳	۹۲/۱۳
درخت تصمیم گرادیان افزایشی	۸۹/۳۵	۹۹/۷۹	۹۱/۳۸
تحلیل سرآیند بسته‌ها (هشدار غلط پایین)	۹۵/۵۲	۹۵/۱۸	۹۵/۴۲
تحلیل سرآیند بسته‌ها (هشدار غلط بالا)	۹۰/۱۲	۹۸/۲۵	۹۱/۸۱
الگوریتم ژنتیک	۹۶/۳۸	۹۶/۲۱	۹۶/۳۳
روش پیشنهادی	۹۷/۷۲	۹۹/۵۳	۹۸/۶۱

۴- نتیجه‌گیری

هدف اصلی در انجام این پژوهش، ارائه یک روش تشخیص نفوذ مبتنی بر رفتار شبکه با استفاده از خوشه‌بندی بهبودیافته مارکوف است. دو هدف اختصاصی برای طراحی روش مورد نظر تعریف شد که شامل استفاده از یک الگوریتم خوشه‌بندی مناسب و همچنین استفاده از معیارهای آماری مناسب جهت تشخیص بهتر ناهنجاری‌ها است. برای نیل به اهداف بالا، مجموعه‌داده مبتنی بر جریان از روی داده‌های DARPA 99 که مبتنی بر بسته است، ایجاد شد. با انتخاب سری زمانی مناسب، مجموعه‌داده ایجادشده به بازه‌های زمانی یکسان تقسیم شد و سپس، گراف مربوط به هر سری زمانی ساخته شده و گراف‌ها با شکل مناسب خوشه‌بندی شدند. با توجه به گوناگونی گراف‌های ایجادشده، الگوریتم خوشه‌بندی مورد استفاده مورد بازبینی و بهینه‌سازی قرار گرفت. به‌منظور ارزیابی و تحلیل نتایج خوشه‌بندی، نتایج حاصل از روش پیشنهادی با هفت روش مطرح دیگر مورد مقایسه قرار گرفت. بررسی نتایج نشان می‌دهد که در روش پیشنهادی، میزان نرخ تشخیص اشتباه به‌اندازه چشم‌گیری کاهش و نرخ دقت روش نسبت به هفت رویکرد دیگر تشخیص نفوذ افزایش پیدا کرده است. دلیل این بهبود عملکرد را می‌توان در عملکرد مناسب الگوریتم خوشه‌بندی بهبودیافته مارکوف دانست که بر روی داده‌های مبتنی بر جریان نتایج دقیق‌تری تولید کرده است. همچنین، تعریف و به‌کارگیری معیارهای مناسب برای تعیین حدود آستانه در کسب نتایج دقیق مؤثر است.

۵- مراجع 5- References

[۱] راستگار، رویا، عیسی‌زاده، آیاز، کریم‌پور، جابر، "تشخیص نفوذ مبتنی بر جریان، براساس خوشه‌بندی گراف پراکندگی ترافیک"، *سیزدهمین کنفرانس بین‌المللی انجمن رمز ایران*، ۱۳۹۵.

[1] R. Rastgar, A. Isazadeh, J. Karimpour, "Flow-based intrusion detection based on traffic distribution graph", In *13th International Conf. on Iranian Cryptography Society*, 2016.

[۲] ضبیحی، مهدیه، وفایی جهان، مجید، "ارائه الگوریتمی بهینه و دقیق مبتنی بر خوشه‌بندی مارکوف برای شناسایی روبات‌های وب"، *هفتمین کنفرانس بین‌المللی انجمن ایرانی تحقیق در عملیات*، ۱۳۹۳.

[2] M. Zabihi, M. Vafaei Jahan, "An optimized accurate algorithm based on Markov clustering for web robots detection", 7th *International Conf. on Iranian Operation Research*, 2014.

[3] S. Anwar, J. M. Zain, M. F. Zolkipli, Z. Inayat, S. Khan, B. Anthony, V. Chang, "From intrusion detection to an intrusion response system: fundamentals, requirements, and future directions", Algorithms, vol. 10(39), 2017.

[4] U.H. Rao, U. Nayak, "Intrusion Detection and Prevention Systems", The InfoSec Handbook. Apress, Berkeley, CA. 2014.

[5] E. Viegas, A. O. Santin, A. Franca, R. Jasinski, V. A. Pedroni and L. S. Oliveira, "Towards an Energy-Efficient Anomaly-Based Intrusion Detection Engine for Embedded Systems"-, *IEEE Transactions on Computers*, vol. 66 (1), pp. 163–177, 2017.

[6] A. Sperotto, "Flow-based intrusion detection", Ph.D. Dissertation, University of Twente, 2010.

[7] L. Hellemons, L. Hendriks, R. Hofstede, A. Sperotto, R. Sadre, and A. Pras, "Sshcure: a flow-based ssh intrusion detection system", *Dependable Networks and Services*, LNCS vol. 7279, pp.86–97, 2012.

[8] M. Iliofotou, P. Pappu, M. Faloutsos, M. Mitzenmacher, S. Singh, and G. Varghese, "Network traffic analysis using traffic dispersion graphs (tdgs): techniques and hardware implementation", UCR Technical Report, 2007.

[9] D. Q. Le, T. Jeong, H. E. Roman, and J. W.-K. Hong, "Traffic dispersion graph based anomaly detection", In *Proc. of the Second Sym. on Information and Communication Technology*, pp.36–41, ACM, 2011.

[10] D. Q. Le, T. Jeong, H. E. Roman, and J. W. Hong, "Trafic dispersion graph based anomaly detection", In *Proc. of the Second Sym. on Information and Communication Technology*, pp. 36–41, ACM, 2011

[11] P. Manandhar and Z. Aung, "Towards practical anomaly-based intrusion detection by outlier mining on tcp packets", *Database and Expert Systems Applications*, LNCS vol. 8645, pp. 164–173, 2014.

[12] H. A. Kholidy, F. Baiardi, "CIDD: A Cloud Intrusion Detection Dataset for Cloud Computing and Masquerade Attacks", 9th *International Conference on Information Technology - New Generations*, Las Vegas, NV, USA, 2012.

[13] R. Lippmann, J.W. Haines, D. J. Fried, J. Korba, and K. Das, "The 1999 darpa off-line intrusion detection evaluation", *Computer networks*, vol. 34(4), pp. 579–595, 2000.

[14] S. M. Dongen, "Graph Clustering by Flow Simulation", PhD Dissertation, University of Utrecht, 2000.

[15] G. P. Guptaa, M. Kularivaa, "A Framework for Fast and Efficient Cyber Security Network Intrusion Detection using Apache Spark", 6*th Int. Conf. on Advances in Computing & Communications*, 2016, 6-8 September 2016. Procedia Computer Science 93, pp. 824 – 831.

فاطمه میرجلیلی مدرک کارشناسی‌ارشد خود را در رشته علوم رایانه گرایش سامانه‌های هوشمند از دانشگاه تبریز دریافت کرده است. زمینه‌های پژوهشی مورد علاقه ایشان الگوریتم‌های پردازش داده‌های بزرگ، داده‌کاوی و یادگیری ماشین است.

نشانی رایانامه ایشان عبارت است از:

f.mirjalili93@ms.tabrizu.ac.ir

جعفر رزم‌آرا دانشیار گروه علوم رایانه دانشگاه تبریز است. ایشان دارای مدرک دکترای علوم رایانه از دانشگاه UTM مالزی بوده و مدرک کارشناسی‌ارشد و کارشناسی خود را در رشته مهندسی کامپیوتر- نرم‌افزار به‌ترتیب از دانشگاه‌های تربیت مدرس و صنعتی اصفهان دریافت کرده است. زمینه‌های پژوهشی مورد علاقه ایشان، کاربرد الگوریتم‌های یادگیری ماشین در طراحی و ساخت سامانه‌های هوشمند، پردازش داده‌های زیستی و بیوانفورماتیک است.

نشانی رایانامه ایشان عبارت است از:

razmara@tabrizu.ac.ir