Journal of Electronical & Cyber Defence
Vol. 9, No. 3, 2021, Serial No. 35

Intrusion Detection in Computer Networks Using Decision Tree and Feature Reduction

A. A. Tajari Siahmarzkooh*

*Assistant Professor, Golestan University, Gorgan, Iran

(Received: 25/11/2020, Accepted: 11/04/2021)

ABSTRACT

Today, the need for anomaly-based intrusion detection systems is felt more than ever due to the emergence of new attacks and the increase in Internet speed. The main criterion for determining the validity of an efficient intrusion detection system is the detection of attacks with high accuracy. In addition to inability of existing systems to manage growing attacks, also they have high rates of positive and negative misdiagnosis. This paper uses the ID3 decision tree features for anomaly-based intrusion detection systems. Two feature selection methods are also used to reduce the amount of used data for the detection and categorization. The KDD Cup99 dataset was used to evaluate the proposed algorithm. The test results show a detection accuracy of 99.89% for the DoS attack and an average accuracy of 94.65% for all attacks using the decision tree, indicating better values than previous tasks.

Keywords: Intrusion Detection, Decision Tree, K-means Clustering, DoS Attack, KDD Cup99 Dataset.

* Corresponding Author Email: a.tajari@gu.ac.ir

نشریه علمی "پدافند الکترونیکی و سایبری"
سال نهم، شماره ۳، پاییز ۱۴۰۰، ص ۱۰۸- ۹۹

علمی- پژوهشی

تشخیص نفوذ در شبکه‌های رایانه‌ای با استفاده از درخت تصمیم و کاهش ویژگی‌ها

علی‌اکبر تجری سیاه‌مرزکوه*
استادیار، گروه علوم کامپیوتر، دانشگاه گلستان، گرگان

(دریافت: ۱۳۹۹/۰۹/۰۵، پذیرش: ۱۴۰۰/۰۱/۲۲)

چکیده

امروزه نیاز به سیستم‌های تشخیص نفوذ مبتنی بر ناهنجاری به‌دلیل ظهور حملات جدید و افزایش سرعت اینترنت بیشتر از قبل احساس می‌شود. معیار اصلی برای تعیین اعتبار یک سیستم تشخیص نفوذ کارآمد، تشخیص حملات با دقّت بالا است. سیستم‌های موجود علاوه بر ناتوانی در مدیریت رو به رشد حملات، دارای نرخ‌های بالای تشخیص مثبت و منفی نادرست نیز می‌باشند. در این مقاله از ویژگی‌های درخت تصمیم ID3 برای سیستم‌های تشخیص نفوذ مبتنی بر ناهنجاری استفاده می‌شود. همچنین از دو روش انتخاب ویژگی برای کاهش میزان داده‌های استفاده شده برای تشخیص و دسته‌بندی استفاده می‌شود. برای ارزیابی الگوریتم پیشنهادی از مجموعه داده KDD Cup99 استفاده شده است. نتایج آزمایش نشان دهنده میزان دقّت تشخیص برای حمله DoS به میزان ۹۹/۸۹٪ و به‌طور میانگین میزان دقّت ۹۴/۶۵٪ برای کلّیه حملات با استفاده از درخت تصمیم است که بیانگر مقادیر بهتر نسبت به کارهای قبلی است.

کلید واژه‌ها: تشخیص نفوذ، درخت تصمیم، خوشه‌بندی k-means، حمله DoS، مجموعه داده KDD Cup99

۱- مقدمه

اینترنت[1] ضروری‌ترین بخش از فعّالیت‌های روزانه ماست که با توجّه به آمارهای جهانی استفاده از اینترنت، تعداد کاربران آن رو به افزایش است. هم‌زمان با این افزایش، تقاضا برای دریافت سرویس‌های آن نیز افزایش می‌یابد، بنابراین داده‌های حسّاس بیشتری در آن مبادله شده و وجود سیستم مطمئن، دقیق و با امنیت بالا برای فراهم کردن اتّصالات ارتباطی و پشتیبانی از داده‌های اشتراکی در سرتاسر اینترنت امری ضروری است. با وجود توسعه گسترده دیواره‌های آتش، ضد ویروس‌ها و سیستم‌های تشخیص نفوذ، فعّالیت‌های مخرّب اینترنتی رو به رشد هستند.

سیستم‌های تشخیص نفوذ جزء حیاتی امنیت رایانه‌ها برای تشخیص حملات در هر رده هستند. تلاش این سیستم‌ها تشخیص حملات به صورت بلادرنگ از طریق مانیتور کردن و تحلیل ترافیک شبکه است که این مورد با جستجو در نشانه‌های حملات یا تفاوت رفتار داده‌ها با داده‌های عادی قابل دستیابی است، امّا این سیستم‌ها در پردازش داده‌های حجیم با شکست مواجه می‌شوند و نمی‌توانند دقّت و حسّاسیت مورد نظر را فراهم نمایند. الگوریتم‌های زیادی برای مقابله با ترافیک رو به رشد، پایگاه داده‌های بزرگ از نشانه‌های حملات، پروفایل‌های رفتاری بزرگ و مشکل تشخیص تفاوت مرز داده‌های عادی و رفتارهای مخرّب ارائه شده است. به‌دلیل وجود تعداد زیاد ویژگی‌ها، دسته‌بندی ترافیک شبکه برای تشخیص رفتار عادی و ناهنجار، یکی از اهداف اصلی سیستم‌های تشخیص نفوذ به‌شمار می‌آید. کشف روابط پیچیده بین ویژگی‌ها کار آسانی نیست، بنابراین تمرکز اصلی در این مقاله انتخاب ویژگی‌های تمییزدهنده بین داده‌ها و توسعه الگوریتم یادگیری احتمالاتی در مقوله نرخ‌های تشخیص بالا و همچنین پردازش سریع در آموزش و آزمایش داده‌هاست.

در این مقاله راهکاری برای بهبود دقّت و سرعت سیستم‌های تشخیص نفوذ با بهره گیری از مزایا و ویژگی‌های درخت تصمیم ارائه شده است. از طرفی زمان و پیچیدگی محاسباتی با به کارگیری روش‌های انتخاب ویژگی و البته با حفظ میزان دقّت تشخیص کاهش یافته است. در این مقاله تئوری احتمالاتی درخت تصمیم برای کار در تشخیص نفوذ مبتنی بر ناهنجاری بهبود داده شده است.

ادامه این مقاله به این شرح تدوین شده است. در بخش دوم، سیستم‌های تشخیص نفوذ و پیش‌زمینه‌ای از درخت‌های تصمیم ارائه شده است. در بخش سوم چند مورد از کارهای مرتبط در این حوزه آورده شده‌اند. در بخش چهارم راهکار و الگوریتم پیشنهادی توصیف شده است. در بخش پنجم جزئیات راهکار پیشنهادی بیان شده است. در بخش ششم معیارهای ارزیابی راهکار پیشنهادی و همچنین نتایج آزمایش به‌همراه مقایسه با راهکارهای پیشین ارائه شده است. در بخش هفتم نتیجه‌گیری به نمایش گذاشته شده است.

*رایانامه نویسنده مسئول: a.tajari@gu.ac.ir

۲- پیش‌زمینه

حملات عملیاتی هستند که تلاش می‌کنند تا یکی از شرایط زیر برای سیستم‌های کامپیوتری را نقض کنند: محرمانگی، صحّت و دسترس‌پذیری. بخش‌های پیش رو پیش‌زمینه‌های مورد نیاز برای فهم مساله را بیان می‌کنند.

۲-۱- سیستم‌های تشخیص نفوذ

سیستم‌های تشخیص نفوذ برای امن نگه داشتن سیستم‌های کامپیوتری استفاده می‌شوند. این سیستم‌ها رفتارها و فعّالیت‌های سیستم‌های کامپیوتری یا شبکه‌ها را مانیتور کرده و آن را برای تشخیص حملات تحلیل می‌کنند. در کل دو نوع سیستم تشخیص نفوذ وجود دارد. سیستم‌های تشخیص نفوذ مبتنی بر میزبان به‌طور مستقیم روی میزبان کار می‌کنند و داده‌های ذخیره شده در آن مانند فرآیندهای در حال اجرا و کاربران داده شده را بررسی می‌کنند. از سوی دیگر سیستم‌های مبتنی بر شبکه، ترافیک مبادله شده بین میزبان‌ها را بررسی کرده و اغلب در نقطه ورود شبکه نصب می‌شوند.

همچنین سیستم‌های تشخیص نفوذ بر اساس روش تشخیص به دو دسته تقسیم‌بندی می‌شوند: تشخیص سوء استفاده و تشخیص ناهنجاری. تشخیص مبتنی بر سوء استفاده یا امضاء متّکی به قوانین موجود برای مقایسه نشانه‌های حمله از پیش تعریف شده با داده‌های جمع‌آوری شده است. مزیت اصلی این روش، نرخ‌های بالای دقّت است امّا امکان شناسایی حملات جدید در آن وجود ندارد. برای رفع این مشکل، پایگاه داده نشانه‌ها باید به‌طور منظّم و به صورت خودکار یا دستی به‌روزرسانی شوند. اسنورت یکی از مشهورترین سیستم‌های تشخیص نفوذ مبتنی بر سوء استفاده است.

سیستم‌های تشخیص نفوذ مبتنی بر ناهنجاری، فعّالیت‌های غیر نرمال را با کشف تفاوت‌های آن‌ها با رفتار عادی سیستم پیدا می‌کنند. تشخیص ناهنجاری به دو نوع تقسیم می‌شود: تشخیص ایستا و پویا. در تشخیص ایستا فرض بر این است که رفتار سیستم هدف تغییر نمی‌کند، بنابراین حملات با انحراف از رفتار ایستای سیستم شناسایی می‌شوند. تشخیص پویا رفتار را در دوره‌های متنوّع بررسی کرده و روند تکاملی برای عملیات در نظر می‌گیرد. مزیت اصلی روش تشخیص ناهنجاری این است که قادر به تشخیص حملات ناشناخته می‌باشد. وجود اختلافات در تعیین مرزها بین رفتارهای نرمال و غیر نرمال منجر به ایجاد نرخ‌های بالای مثبت و منفی نادرست می‌شود.

۲-۲- روش‌های تشخیص نفوذ مبتنی بر ناهنجاری

به‌طور کلّی سیستم‌های تشخیص نفوذ مبتنی بر ناهنجاری بر پایه روش‌های اساسی زیر دسته‌بندی می‌شوند: آماری، مبتنی بر دانش و مبتنی بر یادگیری ماشین. در تشخیص مبتنی بر روش‌های آماری، فعّالیت‌های ترافیک شبکه جمع‌آوری شده و برای محاسبه یک مجموعه از ویژگی‌هایی که اساس رفتار داده‌ها را مشخّص می‌کنند پردازش می‌شوند. به‌عنوان مثالی از ویژگی‌ها می‌توان نرخ اتّصالات و ورودهای ناموفّق را معرّفی کرد. سیستم‌های تشخیص نفوذ آماری به حدّ آستانه و یا پروفایل رفتاری داده‌ها بستگی دارند. این سیستم‌ها محاسبه می‌کنند که چند بار یک رخداد خطرناک در یک دوره مشخّص اتّفاق می‌افتد و اگر یک حدّ آستانه‌ای عبور کرد هشداری تولید می‌شود که این موضوع، نرخ مثبت و منفی نادرست را افزایش می‌دهد.

در روش تشخیص مبتنی بر دانش، یک فرد خبره رخدادها را بر پایه مجموعه‌ای از قوانین به این ترتیب دسته‌بندی می‌کند. ابتدا فرد خبره، ویژگی‌ها و کلاس‌های مجموعه داده‌ی آموزشی را مشخّص می‌کند. سپس قوانین دسته‌بندی را تعریف نموده و در مرحله‌ی بعدی، رخدادهای جدید را بر اساس قوانین دسته‌بندی می‌کند. در پایان، سیستم‌های تشخیص نفوذ مبتنی بر یادگیری طرحی ایجاد می‌کنند که بر اساس آن رخدادهای جدید می‌توانند دسته‌بندی شوند.

۲-۳- درخت تصمیم

درخت تصمیم، یکی از ابزارهای پشتیبانی از مساله تصمیم است که عموماً برای مشخّص کردن روشی که بیشترین احتمال برای رسیدن به هدف دارد استفاده می‌شود. درخت تصمیم سعی می‌کند تا داده‌ها را به گونه‌ای از هم متمایز کند که در هر بخش داده‌های مشابه‌تر قرار گرفته و داده‌های متفاوت‌تر در دسته‌های جداگانه قرار بگیرند. از مزایای درخت تصمیم فهم ساده و کار کردن آن با داده‌های بزرگ است. همچنین از درخت تصمیم می‌توان برای ترکیب با روش‌های دیگر نیز استفاده نمود. در مورد این موضوع در بخش تشخیص نفوذ با استفاده از درخت تصمیم به‌طور دقیق‌تر توضیح داده خواهد شد.

۳- کارهای انجام شده

بسیاری از مطالعات تحقیقاتی انجام شده در حوزه تشخیص نفوذ مبتنی بر ناهنجاری از زمان دنینگ [۱] آغاز شده است. با این وجود بر خلاف حجم زیاد کارهای انجام شده، کاربردهای عملی اندکی از سیستم‌های تشخیص نفوذ مبتنی بر ناهنجاری در محیط عملیاتی گزارش شده است. در سال‌های اخیر، برخی

محقّقان سؤالاتی را در زمینه کاربردی بودن نتایج به‌دست آمده در مقالات مطرح کرده‌اند. گیتس و تیلور [۲] به این نتیجه رسیدند که تنها تعداد اندکی از سیستم‌های تشخیص نفوذ به‌طور گسترده مورد استفاده قرار می‌گیرند. آن‌ها فرضیه‌های موجود در کار دنینگ را سرلوحه‌ی کار خود قرار دادند. به گفته‌ی آنها عدم وجود داده‌های آموزشی و متدولوژی‌های آزمایش مانند تغییرات پیوسته در محتویات، حجم داده‌ها و حملات که ویژگی‌های شبکه را در نظر می‌گیرد دلیل اصلی عدم موفّقیت راهکارهای تشخیص ناهنجاری است.

سامر و پاکسون [۳] مرور جامعی بر تشخیص نفوذ داشتند. استدلال آنها این بود که این حوزه به‌طور قابل ملاحظه‌ای متفاوت از دیگر حوزه‌هایی است که در آنها روش یادگیری ماشین به‌طور مؤثّر استفاده شده است. آنها ادّعا کردند که یادگیری ماشین برای تشخیص شباهت‌ها نسبت به تشخیص نقاط پرت مؤثّرتر است. هزینه‌های بالای مربوط به خطاها استفاده از آن را در محیط عملیاتی غیر ممکن می‌کند. عدم وجود داده‌های به روز و عمومی مانع از ارزیابی و مقایسه درست سیستم‌ها شده است [۴ و ۵]. طبق تحلیل سامر و پاکسون [۳] و پاکسون و فلوید [۶]، محیط دنیای واقعی به‌طور قابل ملاحظه‌ای رفتار متفاوت با داده‌هایی دارد که به‌طور عادی آموزش داده شده‌اند.

قابلیت اطمینان یک سیستم تشخیص نفوذ مبتنی بر ناهنجاری وابسته به مجموعه داده‌ی آموزشی است که به‌طور دقیق آموزش داده شده است. در مورد داده‌های آموزشی باید فرضیه‌های مشخّص و درستی اتّخاذ شود. کانالی و همکاران [۷] مجموعه داده‌ی خود را با جمع‌آوری محتویات سایت‌های مختلف در اینترنت ایجاد کردند؛ آنها داده‌های موجود را با به کارگیری ابزارها و و وارسی داده‌ها برچسب‌گذاری نمودند تا مطمئن شوند محتویات سایت مورد نظر به درستی برچسب‌گذاری شده است. آنها همچنین فرض کردند که اغلب سایت‌هایی که به‌طور مکرّر دیده شده‌اند بی‌خطر هستند و بنابراین توزیع مقادیر ویژگی برای هر کلاس از سایت مورد نظر متفاوت است. مهم‌ترین فرضیه این است که توزیع ویژگی موجود در مجموعه داده استفاده شده برای آموزش مدل‌ها همان توزیع موجود در محیط واقعی را نشان می‌دهد.

علاوه بر این، زمانی‌که یک مجموعه داده در یک محیط کنترل شده به‌دست می‌آید رفتار کاربر به‌طور نرمال از لحاظ آماری تولید مجدّد می‌شود. شیراوی و همکاران [۸] پروفایل کاربران را بر پایه‌ی رفتار کاربران برای هر محیط عملیاتی در هر بازه‌ی زمانی مشاهده شده ایجاد کردند. کندال [۹] با استفاده از تولید مجدّد رفتار کاربران در محیط نیروی هوایی به‌طور آماری، یک مجموعه داده ایجاد کرد. به‌طور کلّی این راهکارها قابلیت ارتقا ندارند و به اشتباه فرض کرده‌اند که ترافیک شبکه قابل تغییر نیست و بر این اساس رفتار کاربر را الگوسازی کردند. بنابراین با روشی که در این مقاله ارائه شده، سعی بر این است تا با بهره‌گیری از برخی راهکارهای انتخاب ویژگی، مجموعه داده مناسبی برای تشخیص نفوذ به‌دست‌آید.

انتخاب ویژگی چند هدفه در بسیاری از فیلدها مانند محدودیت‌های کیفیت سرویس در توزیع محتوای چندرسانه‌ای [۱۰] و تشخیص رقم دست‌نویس[1] [۱۱ و ۱۲] استفاده شده است، با این وجود استفاده از آن برای اهداف تشخیص نفوذ به ندرت اتّفاق افتاده است. دی‌لا هوز و همکاران [۱۳] اندازه‌گیری شباهت بین برچسب‌های پیش‌بینی شده و مجموعه داده‌ای که به درستی برچسب دارد را به عنوان هدف انتخاب کردند که باید بیشترین مقدار را داشته باشد. آنها توانستند دقّت را با کاهش تعداد ویژگی‌ها بهبود ببخشند.

ویگاس و همکاران [۱۴] یک روش انتخاب ویژگی برای پیدا کردن تعادل بین دقّت تشخیص نفوذ و مصرف انرژی سیستم ارائه دادند. آن‌ها با روش پیشنهادی خود توانسته بودند تا ۹۳٪ کاهش مصرف انرژی را با تنها کاهش ۰/۹٪ در دقّت فراهم کنند.

خوان و همکاران [۱۵] یک طرح یادگیری عمیق دو مرحله‌ای را برای تشخیص نفوذ ارائه دادند. مرحله اوّل راهکار پیشنهادی آنها مسئول دسته‌بندی ترافیک شبکه به دو صورت عادی و نرمال با استفاده از یک تابع احتمال است. در مرحله دوم و در واقع مرحله پایانی، از آن به‌عنوان یک ویژگی اضافی برای تشخیص دقیق حملات و داده‌های عادی استفاده می‌شود.

لو و همکاران [۱۶] به منظور بهبود نرخ تشخیص و کاهش نرخ خطا، یک روش جدید تشخیص نفوذ ترکیبی از شبکه عصبی کانولوشن و اصلاح آستانه بر اساس منحنی مشخصه عملکرد[2] ارائه دادند. در این روش از شبکه عصبی کانولوشن به‌عنوان طبقه‌بند استفاده می‌شود و آستانه طبقه‌بند از طریق منحنی اصلاح می‌شود.

آهسان و همکاران [۱۷] برای غلبه بر مشکل نرخ بالای هشدارهای نادرست از الگوریتم حداقل دترمینان کوواریانس[3] و تخمین چگالی هسته[4] استفاده کردند. روش تخمین چگالی از الگوهای داده‌های شبکه پیروی می‌کند و در نتیجه باعث کاهش هشدارهای نادرست می‌شود. از طرفی، استفاده از دترمینان کوواریانس باعث بهبود قابلیت‌های تشخیص برای شناسایی سریع نقاط پرت می‌شود.

[1] Handwritten Digit Recognition
[2] ROC: Receiver Operating Characteristic
[3] MCD: Minimum Covariance Determinant
[4] KDE: Kernel Density Estimation

مصطفی و همکاران [۱۸] یک روش جدید تحلیل منطقه هندسی[1] مبتنی بر تخمین منطقه ذوزنقه‌ای[2] برای مشاهدات صورت گرفته در ترافیک شبکه ارائه دادند. همچنین گاتوالت و همکاران [۱۹] یک روش جدید انتخاب ویژگی برای سیستم‌های تشخیص نفوذ مبتنی بر همبستگی چند متغیره ارائه دادند. نتایج به دست آمده از این روش نشان دهنده بهبودها در انتخاب ویژگی‌های مؤثرتردر مجموعه داده NSL-KDD است.

۴- راهکار پیشنهادی

جریان کلّی راهکار پیشنهادی بر پایه‌ی الگوریتم درخت تصمیم و اصول داده‌کاوی برای تشخیص نفوذ بنا شده است. به دلیل کیفیت پیش‌پردازش‌های مورد استفاده، الگوریتم احتمالاتی پیشنهادی به یک نامزد کامل برای تشخیص نفوذ تبدیل شده است به گونه‌ای که در میان حجم عظیم داده‌ها قادر به شناسایی رفتارهای عادی و غیر عادی خواهد بود.

در این مقاله از الگوریتم درخت تصمیم برای سیستم تشخیص نفوذ استفاده شده است. الگوریتم درخت تصمیم شامل سه مرحله است که در شکل (۱) نشان داده شده است. استخراج و انتخاب ویژگی‌های ورودی با دسته‌بندی و روش‌های احتمالاتی تحلیل همبستگی صورت گرفته است. در مرحله پیش‌پردازش داده‌ها، داده‌ها برای دسته‌بندی و حذف ویژگی‌های بی‌فایده جهت آموزش و آزمایش آماده می‌شوند. الگوریتم پیشنهادی داده‌ها را در حافظه ذخیره کرده و همان‌طور که در بخش ۵ شرح داده شده رکوردها را یکی پس از دیگری می‌خواند. در بخش بعدی، مجموعه داده آموزش و آزمایش، انتخاب ویژگی و پیش‌پردازش داده‌ها شرح داده شده است.

۴-۱- مجموعه داده KDD Cup99

مجموعه داده KDD Cup99 مجموعه داده استاندارد آموزش و آزمایش سیستم‌های تشخیص نفوذ مبتنی بر ناهنجاری است. این مجموعه داده از مجموعه داده شبکه دارپا برگرفته شده که توسّط آزمایشگاه‌های فناوری در مؤسّسه ماساچوست جمع‌آوری شده است. مجموعه داده دارپا مربوط به شبکه نظامی نیروی هوایی ایالات متّحده است که مورد حملات متعدّد قرار گرفته است. هر رکورد در مجموعه داده KDD Cup99 نشان‌دهنده ۴۱ ویژگی برای یک اتّصال به همراه یک برچسب کلاس است که این برچسب نشان می‌دهد آیا این اتّصال عادی است یا اینکه متعلّق به یکی از ۲۲ نوع حمله مختلف است. مجموعه داده آزمایش شامل ۱۴ نوع حمله بیشتر است که در مجموعه داده آزمایش وجود ندارد و بیانگر حمله ناشناخته است.

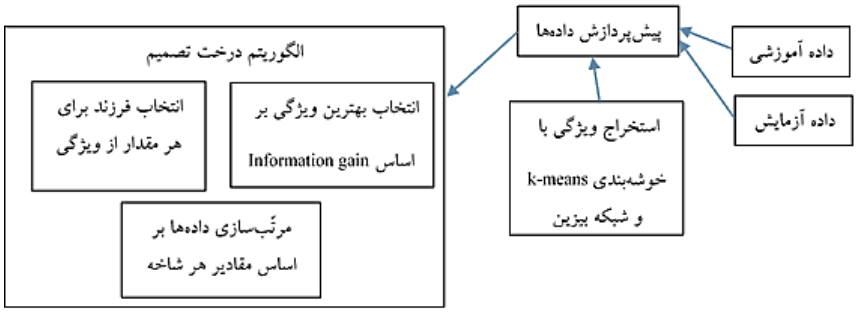
این ۴۱ ویژگی اتّصال شامل ۳۸ ویژگی عددی و سه ویژگی غیر عددی (نمادین) است. این ویژگی‌ها در سه گروه زیر قرار می‌گیرند: ۹ ویژگی اتّصالات پایه مانند دوره، نوع پروتکل، سرویس، پرچم و بایت‌های منبع و مقصد؛ ۱۳ ویژگی توصیف‌کننده اتّصالات توسّط آگاهی از حوزه آن‌ها است مانند ورودهای ناموفّق، تلاش برای دستیابی به ریشه، تلاش کابر فوق‌العاده و ایجاد فایل؛ ۱۳ ویژگی ترافیک که توصیف‌کننده اتّصالات دوطرفه هستند مانند تعداد اتّصالات به یک میزبان مشخّص، اتّصال به یک سرویس مشخّص و نرخ اتّصالاتی که خطای همزمانی دارند.

جدول (۱): توزیع کلاس حملات در مجموعه داده KDD Cup99

نوع حمله	تعداد داده آموزشی	تعداد داده آزمایش	درصد در داده آموزشی	درصد در داده آزمایش
عادی	۹۷۲۷۷	۶۰۵۹۲	۱۹/۶۹	۱۹/۴۸
Probing	۴۱۰۷	۴۱۶۶	۰/۸۳	۱/۳۴
DoS	۳۸۹۱۴۵۸	۲۲۹۸۵۳	۷۹/۲۴	۷۳/۹۰
U2R	۵۲	۷۰	۰/۰۱	۰/۰۲
R2L	۱۱۲۶	۱۶۳۴۴	۰/۲۳	۵/۲۵
مجموع	۴۹۴۰۲۰	۳۱۱۰۲۸	۱۰۰	۱۰۰

خطاهای موجود در این مجموعه داده در یکی از این ۴ کلاس قرار می‌گیرد. حمله عدم سرویس[3] سعی می‌کند تا منابع کامپیوتر را به وسیله سرویس‌های طغیان‌گر غیر قابل دسترس نماید. حملات اکتشافی[4] تلاش می‌کنند تا شبکه را برای جمع‌آوری اطّلاعات جهت یافتن آسیب‌پذیری‌ها پویش نماید. حمله کاربر به ریشه[5] ابتدا به‌عنوان کاربر عادی به یک سیستم دسترسی پیدا می‌کند و سپس سعی می‌کند تا با استخراج ضعف‌های سیستم به حریم خصوصی ریشه دست پیدا کند [۲۰]. در نهایت حمله از راه دور[6] زمانی رخ می‌دهد که حمله‌کننده، بسته‌ها را به یک ماشین از راه دور ارسال می‌کند تا نقص‌ها را پیدا کرده و به آن‌ها دسترسی محلّی پیدا کند [۲۱].

[1] GAA: Geometric Area Analysis
[2] TAE: Trapezoidal Area Estimation
[3] DoS: Denial of Service
[4] Probing Attacks
[5] User to Root (U2R)
[6] Remote to Local (R2L)



شکل (۱): طرح دسته‌بندی درخت تصمیم

مجموعه داده آموزشی حدوداً شامل ۵ میلیون رکورد است و همچنین یک نسخه کوچکتر شامل ۱۰٪ داده‌های آن نیز موجود است. جدول (۱) تعداد و درصد هر کلاس از حملات و داده‌های عادی را در داده‌ها نشان می‌دهد. اغلب رکوردها متعلّق به کلاس‌های حمله DoS و عادی هستند به گونه‌ای که بعداً خواهیم دید نتایج نرخ‌های تشخیص پایین به‌دلیل عدم وجود داده‌های آموزشی کافی برای حملات U2R و R2L صورت می‌گیرد.

۴-۲- استخراج و انتخاب ویژگی‌ها

با داشتن یک مجموعه داده بزرگ لازم است تعداد ویژگی‌ها در راستای تشخیص به موقع حملات کاهش یابد. کاهش ویژگی‌ها می‌تواند به یکی از طرق مختلف مانند فیلتر کردن داده‌ها برای کاهش حجم ترافیک به‌کار گرفته شده توسّط سیستم تشخیص نفوذ صورت گیرد که ممکن است منجر به دور انداختن داده‌های مفید شود. از خوشه‌بندی برای استخراج الگوها در ویژگی‌ها استفاده شده که می‌تواند برای پیدا کردن زیرمجموعه حدّاقلی از ویژگی‌های وابسته به هم مفید باشد.

در این مقاله هر دو راهکار خوشه‌بندی k-means و شبکه بیزین برای انتخاب ویژگی‌های مهم‌تر اعمال می‌شوند.

۴-۲-۱- الگوریتم خوشه‌بندی K-means

ساده‌ترین و بهترین الگوریتم خوشه‌بندی که بعد از محاسبه فاصله هر یک از داده‌ها با چند نقطه مرکزی به‌عنوان سرخوشه، میزان احتمال تعلّق آنها به سرخوشه‌ها محاسبه می‌شود [۲۲]. این روند به‌صورت تکراری و در چند مرحله تکرار می‌شود تا تمام داده‌های قرار گرفته در یک خوشه به یکدیگر نسبت به سایر داده‌ها نزدیکتر باشند. به‌عنوان مثال، این الگوریتم داده‌ها را با توجه به ویژگی سرویس تجزیه می‌کند و چنانچه مقداری برای ویژگی سرویس در دسترس نباشد باید از ویژگی دیگری مانند پرچم به‌عنوان جایگزین استفاده شود. همان‌طور که در جدول (۲) نشان داده شده است این الگوریتم منجر به کاهش تعداد ۴۱ ویژگی به ۱۰ ویژگی شده است.

۴-۲-۲- الگوریتم شبکه بیزین

شبکه بیزین یک طرح گرافیکی برای نمایش ارتباط احتمالی بین یک مجموعه از متغیّرهاست. این شبکه یک گراف جهت‌دار بدون دور (N,A) است که N متغیر حوزه و A کمان بین گره‌هاست که وابستگی بین گره‌ها یا متغیرها را نشان می‌دهد. به‌طور کلّی یک دسته‌بند شبکه بیزین به یک مجموعه از متغیّرهایی دست پیدا می‌کند که روی گره کلاس مؤثّرتر هستند. بنابراین با استفاده از این شبکه ویژگی‌های مؤثّرتر انتخاب شده و سایر ویژگی‌ها حذف می‌شوند [۲۳]. با استفاده از این روش، برخی از ویژگی‌های مؤثّرتر در تعیین کلاس انتخاب شده‌اند که در جدول (۲) نشان داده شده‌اند. این ویژگی‌ها در سه دسته قرار می‌گیرند. اوّلین دسته همان ویژگی‌های میزبان است که اتّصال با یک میزبان مقصد مشخّص را در آخرین اتّصالات بررسی می‌کند که شامل آمار مرتبط با رفتار پروتکل مانند بایت منبع، بایت مقصد و سرویس است. دسته دوم ویژگی‌های سرویس است که تنها اتّصالاتی را بررسی می‌کند که یک سرویس مشخّص را در اتّصالات اخیر دارند. در نهایت سومین دسته ویژگی‌های ترافیک مبتنی بر میزبان است که اتّصالات با مقصد یکسان در ۱۰۰ اتّصال قبلی را بررسی می‌کند. نکته‌ی قابل ذکر اینکه ۸ ویژگی از ۱۰ ویژگی انتخاب شده توسّط خوشه‌بندی در ویژگی‌های انتخاب شده در درخت شبکه بیزین مشاهده می‌شوند.

جدول (۲): ویژگی‌های انتخاب شده با استفاده از خوشه‌بندی k-means و شبکه بیزین

Count	Srv_count	Logged_In	ویژگی‌های خوشه‌بندی k-means
Land	Wrong_ fragment	Src_byte	
Protocol_type	Serror_rate	Is_guest_Login	
Serror_rate	Num_failed_logins	Wrong_fragment	شبکه بیزین
Count	Root_shell	Diff_srv_rate	
Duration	Srv_count	Dst_host_count	
Logged_in	Src_byte	Src_byte	

۴-۳- پیش‌پردازش داده‌ها

قبل از آموزش و آزمایش، داده‌ها پردازش می‌شوند تا مشخّص شود هر ویژگی و مقادیر آن به درستی از مقادیر نمادین به مقادیر عددی نگاشت شده‌اند. ویژگی سرویس مقادیر نمادین دارد (مانند http، smtp و غیره که همگی به اعداد صحیح مثبت نگاشت شده‌اند). در حالت برچسب کلاس، تمامی حمله‌ها به یکی از ۵ مقدار، ۱ برای عادی، ۲ برای probe، ۳ برای DoS، ۴ برای U2R و ۵ برای R2L نگاشت شده‌اند.

۵- تشخیص نفوذ با الگوریتم درخت تصمیم

همان‌طور که در الگوریتم شکل (۲) نشان داده شده است زمانی که رکوردهای پیش‌پردازش شده به الگوریتم درخت تصمیم داده می‌شوند ابتدا درخت به‌صورت بازگشتی با روش تقسیم و حل از بالا به پایین ساخته می‌شود. در آغاز تمام نمونه‌های آموزشی در ریشه هستند. برای ویژگی‌های پیوسته، ابتدا گسسته‌سازی روی آنها صورت می‌گیرد و داده‌ها به ورت بازگشتی و بر اساس ویژگی‌های انتخاب شده دسته‌بندی می‌شوند. ویژگی‌های آزمون نیز بر اساس شاخص‌های آماری انتخاب می‌شوند. برای توقّف دسته‌بندی، شرایط زیر را در نظر می‌گیریم: اوّل اینکه هیچ ویژگی دیگری برای دسته‌بندی بیشتر باقی نمانده باشد که در این صورت برچسب اغلب نمونه‌ها برای دسته‌بندی به کار گرفته می‌شوند و دوم اینکه همه‌ی نمونه‌ها برای یک گره داده شده متعلّق به یک کلاس باشند.

استنتاج بالا به پایین صورت گرفته در درخت تصمیم به این گونه است. ابتدا با فرض A به عنوان بهترین ویژگی تصمیم برای گره بعدی، برای هر مقدار از این ویژگی، فرزند جدیدی ایجاد می‌شود. در مرحله بعدی با توجه به مقدار ویژگی هر شاخه، داده‌های آموزشی موجود مرتّب‌سازی می‌شوند. در پایان اگر تمامی نمونه‌های آموزشی دسته‌بندی شده باشند کار تمام می‌شود در غیر این صورت برای گره‌های جدید باید الگوریتم از ابتدا تکرار شود.

در صورتی‌که p_i احتمال تعلّق رکورد دلخواه در مجموعه داده D به کلاس C_i باشد در این صورت این مؤلفه برابر $|C_{i,D}| \div |D|$ خواهد بود و معیار انتخاب ویژگی تصمیم بر اساس میزان اطّلاعات مورد انتظار برای دسته‌بندی یک رکورد در مجموعه داده که از رابطه (۱) محاسبه می‌شود به‌دست می‌آید.

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2 p_i \qquad (۱)$$

همچنین اطّلاعات مورد نیاز برای دسته‌بندی مجموعه داده پس از انتخاب ویژگی A برای تقسیم D به n دسته از رابطه (۲) به‌دست می‌آید.

$$Info_A(D) = \sum_{j=1}^{n} \frac{|D_j|}{|D|} * Info(D_j) \qquad (۲)$$

در نهایت information gain به دست آمده به وسیله انشعاب روی ویژگی از رابطه (۳) حاصل می‌شود.

$$Gain(A) = Info(D) - Info_A(D) \qquad (۳)$$

آن ویژگی که بالاترین مقدار information gain را دارد به عنوان ویژگی تصمیم انتخاب می‌شود. انتخاب ویژگی‌ها تا جایی ادامه می‌یابد که یکی از دو شرط خاتمه الگوریتم ذکر شده تحقّق یابد.

Algorithm 1. Decision Tree Algorithm

Inputs: training and testing dataset

Output: the best solution result

For all training records

1: Select the best feature (A) based on **highest information gain**

2: Create a **child** for each value of A

3: Sort training data using value of each branch

4: If all of data is classified **finish**

5: If not **repeat 1**.

شکل (۲): الگوریتم درخت تصمیم [۲۲]

۶- نتایج آزمایش

در این بخش معیارهای کارآیی استفاده شده برای ارزیابی راهکار پیشنهادی بیان شده و نتایج آزمایش الگوریتم پیشنهادی به همراه مقایسه با سایر کارهای مرتبط انجام شده تحلیل و بررسی می‌شوند.

۶-۱- معیارهای کارآیی

برای ارزیابی کارآیی الگوریتم پیشنهادی از شناخته شده‌ترین معیارها استفاده می‌شود: نرخ مثبت صحیح (TPR)، نرخ منفی صحیح (TNR)، نرخ مثبت غلط (FPR) و نرخ دقّت (Accuracy).

$$TPR = \frac{TP}{FN + TP} \qquad (۴)$$

$$TNR = \frac{TN}{TN + TP} \qquad (۵)$$

$$FPR = \frac{TP}{TN + FP} \qquad (۶)$$

$$\text{Accuracy rate} = \frac{TN + TP}{TN + TP + FN + FP} \qquad (۷)$$

تشخیص مثبت صحیح، تعداد حملاتی است که به درستی به عنوان حمله دسته‌بندی شده‌اند، منفی صحیح، تعداد نمونه‌های

عادی است که به درستی برچسب‌گذاری نرمال به آنها داده شده است، مثبت نادرست، تعداد نمونه‌های عادی است که به اشتباه به‌عنوان حمله شناسایی شده‌اند و منفی نادرست تعداد حملاتی است که به‌طور اشتباه نرمال دسته‌بندی شده‌اند.

۶-۲- محیط شبیه‌سازی

الگوریتم پیشنهادی با استفاده از دو مجموعه ویژگی انتخاب شده توسط خوشه‌بندی k-means و درخت شبکه بیزین ارزیابی می‌شود. این الگوریتم به زبان C++ نوشته شده و آزمایش روی یک پردازنده Core i7، ۲/۸ گیگاهرتز با ۴ گیگابایت حافظه اجرا شده است. مجموعه داده KDD Cup 99 برای اندازه‌گیری کارآیی الگوریتم پیشنهادی مورد استفاده قرار گرفته است. مقادیر اوّلیه بر اساس نسخه‌های مختلف آزمایش انتخاب شده است.

۶-۳- نتایج آزمایش با بهره‌گیری از الگوریتم درخت تصمیم و انتخاب ویژگی‌های مهم‌تر

جدول‌های (۳ و ۴) نتایج آزمایش را به‌ترتیب برای حملات شناخته شده و ناشناخته نشان می‌دهد. الگوریتم پیشنهادی تمامی حملات را به‌جز "phf" و "Loadmodule" شناسایی می‌کند. اغلب حملاتی که تشخیص داده نشده‌اند مربوط به کلاس‌های R2L و U2R هستند؛ حملاتی مانند "mailbomb" و "sendmail" اسپم‌هایی هستند که در لایه شبکه عمل می‌کنند و شناسایی آنها به‌دلیل ویژگی‌های اتّصالاتی که دارند کار پیچیده‌ای است. برای تشخیص دقیق آن‌ها از مابقی ترافیک به ویژگی‌های لایه‌ی کاربردی و تحلیل آنها نیاز است.

جدول (۵) نرخ‌های تشخیص برای کلاس حملات شناخته شده و ناشناخته را نشان می‌دهد. همان‌طور که مشاهده می‌شود الگوریتم پیشنهادی بالاترین دقّت تشخیص یعنی ۹۹/۸۹٪ و ۹۷/۷۲٪ را به‌ترتیب برای کلاس حملات شناخته شده DoS و Probe دارد. بیشترین درصد رکوردهای موجود در مجموعه داده آزمایش مربوط به این دو حمله است. با این وجود الگوریتم پیشنهادی نرخ تشخیص چندان بالایی برای U2R و R2L ندارد که دلیل آن تعداد کم کلاس داده‌های این نوع حملات در مجموعه داده آزمایش است.

برای حملات ناشناخته، نرخ تشخیص حمله DoS برابر ۳۲/۵۴٪ است که مقدار کمی است زیرا در مجموعه داده آزمایشی، تعداد ۵۰۰۰ رکورد mailbox از ۶۵۵۵ رکورد یعنی حدود ۷۶/۳٪ وجود دارد. همان‌طور که قبل‌تر گفته شد حملات mailbox در لایه شبکه شناسایی نمی‌شوند. با این حال، نرخ‌های تشخیص برای probe و U2R در مقایسه با حملات شناخته شده بیشتر است امّا نرخ تشخیص حمله R2L نسبتاً کم است.

جدول (۶) نشان می‌دهد که ۶۰۰۱۵ رکورد از ۶۰۵۹۲ رکورد واقعاً عادی به عنوان عادی شناسایی شده‌اند که نشان دهنده ۹۹/۰۴٪ نرخ تشخیص و ۰/۹۶٪ نرخ هشدار نادرست است. سطر آخر بیان کننده میانگین نرخ دقّت ۹۴/۶۵٪ برای تمامی کلاس‌های موجود است.

جدول (۷) مقایسه الگوریتم پیشنهادی را با راهکارهایی مانند k همسایگی، درخت C4.5، ماشین بردار پشتیبان، ژنتیک فازی چند هدفه، بیزین ساده و چند مقاله‌ی دیگر نشان می‌دهد. اگرچه به دلیل وجود ویژگی‌های متغیّر در الگوریتم‌های مختلف، این دقیق‌ترین مقایسه نیست امّا باعث می‌شود کارآیی نسبی بیشتر الگوریتم پیشنهادی را مدّ نظر قرار دهیم.

جدول (۳): نتایج آزمایش برای حملات شناخته شده

نام کلاس	نام حمله	کل رکوردها	تشخیص داده شده	نرخ تشخیص (٪)
DoS	Back	۱۰۹۸	۱۰۹۶	۹۹/۸۱
	Land	۹	۹	۱۰۰
	Neptune	۵۸۰۰۱	۵۷۸۴۹	۹۹/۷۳
	Pod	۸۷	۸۷	۱۰۰
	Smurf	۱۶۴۰۹۱	۱۶۴۰۱۲	۹۹/۹۵
	Teardrop	۱۲	۱۰	۸۳/۳۳
Probe	Ipsweep	۳۰۶	۳۰۴	۹۹/۳۴
	Nmap	۸۴	۸۳	۹۸/۸۰
	Portsweep	۳۵۴	۳۴۷	۹۸/۰۲
	Satan	۱۶۳۳	۱۵۸۹	۹۷/۳۰
R2L	ftp_write	۳	۲	۶۶/۶۶
	Guess_passwd	۴۳۶۷	۲۶۵۱	۶۰/۷۰
	Imap	۱	۱	۱۰۰
	Multihop	۱۸	۱۲	۶۶/۶۶
	Phf	۲	۰	۰
	Warezmaster	۱۶۰۲	۱۴۰۳	۸۷/۵۷
U2R	Buffer_overflow	۲۲	۱۸	۸۱/۸۱
	Loadmodule	۲	۰	۰
	Perl	۲	۲	۱۰۰
	Rootkit	۱۳	۸	۶۱/۵۳

جدول (۴): نتایج آزمایش برای حملات ناشناخته

نام کلاس	نام حمله	کل رکوردها	تشخیص داده شده	نرخ تشخیص (٪)
DoS	Apache2	۷۹۴	۶۴۴	۸۱/۱۰
	Mailbomb	۵۰۰۰	۷۴۳	۱۴/۸۶
	Processtable	۷۵۹	۷۴۴	۹۸/۰۲
	Udpstorm	۲	۲	۱۰۰
Probe	Mscan	۱۰۵۳	۹۹۸	۹۴/۷۷
	Saint	۷۳۶	۷۳۲	۹۴/۴۵
R2L	Httptunnel	۱۵۸	۱۳۹	۸۷/۹۷
	Named	۱۷	۹	۱۷/۵۲
	Sendmail	۱۷	۲	۱۱/۷۶
	Snmpgetattack	۷۷۴۱	۴۳۱	۵/۵۶
	Snmpguess	۲۴۰۶	۱۸۴۰	۷۶/۴۷
	Worm	۲	۲	۱۰۰
	Xclock	۹	۶	۶۶/۶۶
	Xsnoop	۴	۲	۵۰/۰
U2R	Ps	۱۶	۱۵	۹۳/۷۵
	Sqlattack	۲	۲	۱۰۰
	Xterm	۱۳	۱۱	۸۴/۶۱

جدول (۵): نرخ‌های تشخیص برای کلاس حملات شناخته شده و ناشناخته

	کلاس حمله	کل رکوردها	تشخیص داده شده	نرخ تشخیص (٪)
حملات شناخته شده	DoS	۲۲۳۲۹۸	۲۲۳۰۶۳	۹۹/۸۹
	Probe	۲۳۷۷	۲۳۲۳	۹۷/۷۲
	U2R	۳۹	۲۸	۷۱/۷۹
	R2L	۵۹۹۳	۲۶۶۶	۴۴/۴۸
نرخ تشخیص کل				۹۸/۴۳
حملات ناشناخته	DoS	۶۵۵۵	۲۱۳۳	۳۲/۵۴
	Probe	۱۷۸۹	۱۷۳۰	۹۶/۷۰
	U2R	۳۱	۲۸	۹۰/۳۲
	R2L	۱۰۳۵۴	۲۴۳۱	۲۳/۴۷
نرخ تشخیص کل				۳۳/۷۵

جدول (۶): خلاصه نرخ‌های تشخیص با استفاده از الگوریتم درخت تصمیم

نام کلاس	تعداد کل	تشخیص داده شده	نرخ تشخیص (٪)
DoS	۲۲۹۸۵۳	۲۲۵۱۹۶	۹۷/۹۷
Probe	۴۱۶۶	۴۰۵۳	۹۷/۵۲
U2R	۷۰	۵۶	۸۰/۰
R2L	۱۶۳۴۷	۵۰۹۷	۳۱/۱۸
نرمال	۶۰۵۹۳	۶۰۰۱۵	۹۹/۰۴
مجموع	۳۱۱۰۲۹	۲۹۴۴۱۷	۹۴/۶۵

همان‌طور که در جدول (۷) نشان داده شده کـلاس حمـلات R2L و U2R در تمامی راهکارهای تشخیص مقادیر نسبتاً پـایینی دارد که یافته‌های ما را تأیید می‌کند. نرخ‌های کم در تشخیص به این دلیل اسـت کـه حمـلات R2L و U2R رفتـار ترتیبـی ماننـد حملات DoS و probe ندارند زیرا این دو حمله‌ی آخر در یک بازه زمانی کوتاه تعداد اتّصالات زیادی دارند. دلیـل دیگـر بـرای نـرخ تشخیص کم، درصد کم حضور این دو نوع حمله در مجموعه داده آموزشی است که به ترتیب شامل ۱۱۲۶ رکـورد یعنـی ۰/۲۲٪ و ۵۲ رکورد یعنی ۰/۰۱٪ می‌شود.

برای حمله DoS، الگوریتم پیشنهادی بالاترین دقّـت را دارد. برای حملات probe نیز مطلقاً این موضوع صحّت دارد. در نهایـت برای داده‌های عادی نیز الگوریتم پیشنهادی از تمامی الگوریتم‌ها به‌جز cup_winner [۲۴] کارآیی بیشتری دارد.

بدیهی است با در نظر گـرفتن معیارهـای کلیـدی در بـرآورد دقّـت سیسـتم‌هـای تشـخیص نفـوذ و همچنـین آزمـایش روش پیشنهادی بر روی مجموعه داده‌های استاندارد موجـود مـی‌تـوان این نتیجه‌گیری را ارائه نمود: در ارزیابی یـک سیسـتم تشـخیص نفوذ، هر چه مقادیر بیشـتری در معیارهـای تشـخیص درسـت و مقادیر کمتری در معیارهای تشخیص نادرست حملات بـه دسـت آید آن سیستم از کارآیی بیشتری برخوردار است و قابل اعتمادتر است. بنابراین تمرکز اصلی صورت گرفته در ایـن مقالـه، بررسـی پارامترهای کلیدی و اثرگـذار در تشـخیص حمـلات و بـه دسـت آوردن مقدار پارامترهای مـؤثر در تعیـین میـزان دقّـت سیسـتم تشخیص نفوذ است که نتـایج ان در جـداول (۶- ۳) آورده شـده است که به جز موارد خاص ذکر شده که قابل تشخیص دقیق هم نیستند نتایج مناسب و درخوری به دست آمده است. در بخش ۷ نتیجه‌گیری از کار صورت گرفته در این مقاله به عمل آمده است.

جدول (۷). مقایسه راهکار پیشنهادی با دیگر روش‌های موجود

نام کلاس	روش پیشنهادی	یادگیری عمیق در دو گام (TSDL) [۱۵]	ماشین بردار پشتیبان بهینه [۱۶]	KDE-CL تخمین چگالی هسته [۱۷]	درخت C4.5 [۲۵]	K نزدیکترین همسایگی [۲۵]	تحلیل ناحیه جغرافیایی GAA-ADS [۱۹]	همبستگی چند متغیره [۲۵]	تشخیص نفوذ فازی ژنتیک چند هدفه [۱۸]	Cup-winner [۲۴]	GALC [۲۶]	بیزین ساده [۲۵]
نرمال	۹۹/۰۴	۹۳/۲۴	۹۴/۴۴	۹۷/۴۸	۹۸/۳۸	۹۵/۸۹	۹۰/۵۳	۹۴/۲۹	۹۸/۳۶	**۹۹/۵۰**	۹۶/۱۴	۵۵/۴۷
DoS	**۹۷/۹۷**	۷۷/۱۴	۹۶/۸۵	۹۵/۳۲	۹۶/۹۹	۹۷/۰۰	۹۱/۴۹	۹۵/۶۳	۹۷/۲۰	۹۷/۱۰	۹۶/۶۸	۸۲/۷۵
Probe	**۹۷/۵۲**	۹۳/۳۲	۹۶/۷۴	۹۶/۱۹	۸۱/۸۸	۸۱/۶۱	۸۲/۵۲	۹۵/۴۸	۸۸/۶۰	۸۳/۳۰	۸۵/۷۷	۹۰/۴۵
U2R	**۸۰/۰**	۷۷/۰۴	۷۵/۸۳	۷۴/۲۲	۱۴/۴۷	۱۴/۹۱	۲۳/۱۱	۷۶/۵۵	۱۵/۷۹	۱۳/۲۰	۷۵/۷۱	۱۳/۱۶
R2L	۳۱/۱۸	۵۴/۲۲	۵۸/۴۳	۵۵/۴۶	۱/۴۵	۶۰/۹۰	۶۳/۹۶	۶۲/۲۸	۱۱/۰۱	۸/۴۰	۳۰/۳	**۶۲/۷۴**
دقّت	**۹۴/۶۵**	۸۹/۷۱	۹۱/۴۰	۹۱/۰۱	۹۲/۰۲	۹۱/۸۳	۹۲/۸۰	۹۳/۲۲	۹۲/۷۷	۹۲/۷۱	۹۲/۹۴	۷۶/۴۵

۷- نتیجه‌گیری

امنیت شبکه‌های کامپیوتری به‌دلیل افزایش سرعت اینترنت و تعداد سرویس‌های ضروری موجود در شبکه‌های کامپیوتری روز به روز اهمّیت بیشتری پیدا می‌کند. سیستم‌های تشخیص و پیشگیری از نفوذ یک جزء اصلی از معماری کامپیوتر به‌حساب می‌آیند. در این مقاله راهکار جدیدی مبتنی بر درخت تصمیم برای تشخیص نفوذ ارائه شد. موضوع مورد استفاده دیگر بهره‌گیری از روش انتخاب ویژگی بود که دقّت تشخیص را افزایش و پیچیدگی محاسباتی را کاهش داد. راهکار پیشنهادی روی مجموعه داده KDD Cup 99 مورد آزمایش قرار گرفت. نرخ بالای دقّت و همچنین برتری روش پیشنهادی نسبت به سایر راهکارهای ارائه شده توسّط دیگر محقّقان نشان می‌دهد ترکیب روش‌های استفاده شده در این مقاله می‌تواند ابتکار مناسبی در جهت بهبود عملکرد سیستم‌های تشخیص نفوذ باشد.

* این پژوهش با استفاده از اعتبارات دانشگاه گلستان در قالب طرح تحقیقاتی شماره ۹۹۲۰۳۱ انجام گردیده است.

۸- مراجع

[1] D.E. Denning, “An intrusion detection model,” IEEE Symposium on Security and Privacy, vol. 13, pp. 222-232, 1997.

[2] C. Gates, C. Taylor, “Challenging the anomaly detection paradigm: A provocative discussion,” Proceedings of 2006 Workshop, New Security Paradigms, pp. 21-29, 2007.

[3] R. Sommer, V. Paxson, “Outside the closed world: On using machine learning for network intrusion detection,” Proceedings of IEEE Symposium on Security and Privacy, pp.305-316, 2010.

[4] J. Peng, K.K.R. Choo, H. Ashman, “User profiling in intrusion detection: A review,” Journal of Network and Computer Applications, vol. 72, pp.14-27, 2016.

[5] A.I. Abubakar, H. Chiroma, S.A. Muaz, L.B. Ila, “A review of the advances in cyber security benchmark datasets for evaluating data-driven based intrusion detection systems,” Procedia Computer Science, vol. 62, pp. 221-227, 2015.

[6] V. Paxson, S. Floyd, “Wide-area traffic: The failure of Poisson modeling,” IEEE/ACM Transactions on Networking, vol. 3, pp. 226-244, 1995.

[7] D. Canali, M. Cova, G. Vigna, C. Kruegel, “Prophiler: A fast filter for the large -scale detection of malicious Web pages categories and subject descriptors,” Proceedings of. International World Wide Web Conference, pp.197-206, 2017.

[8] A. Shiravi, H. Shiravi, M. Tavallaee, A. Ghorbani, “Toward developing a systematic approach to generate benchmark datasets for intrusion detection,” Computers & Security, vol. 31, pp. 357-354, 2012.

[9] M. Tavallaee, E. Bagheri, W. Lu, A. a. Ghorbani, “A detailed analysis of the KDD CUP 99 data set,” IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA), pp. 1-6, 2009.

[10] U. Shaukat, Z. Anwar, “A fast and scalable technique for constructing multicast routing trees with optimized quality of service using a firefly based genetic algorithm,” Multimedia Tools and Applications, vol. 75, pp. 2275-2301, 2016.

[11] L. S. Oliveira, R. Sabourin, F. Bortolozzi, C.Y. Suen, “A methodology feature selection using multi-objective genetic algorithms for handwritten digit string recognition,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 17, pp. 903-929, 2003.

[12] Z. Fei, B. Li, S. Yang, C. Xing, H. Chen, L. Hanzo, “A survey of multi-objective optimization in wireless sensor networks: Metrics Algorithms and Open Problems,” in: IEEE Communications Surveys & Tutorials, vol. 19, pp. 550-586, 2017.

[13] E. De, A. Ortiz, A. Martinez-Alvarez, “Feature selection by multi-objective optimization: Application to network anomaly detection by hierarchical self-organizing maps,” Knowledge-based Systems, vol. 71, pp. 322-338, 2014.

[14] E. Viegas, A. Santin, A. Franca, R. Jasinksi, V. Pedroni, L. Oliveira, "Towards an energy-efficient anomaly-based intrusion detection engine for embedded systems," IEEE Transactions on Computers, vol. 66, pp. 163-177, 2017.

[15] F. A. Khan, A. Gumaei, A. Derhab, A. Hussain, "A novel two-stage deep learning model for efficient network intrusion detection," IEEE Access, vol. 7, pp. 30373–30385, 2019.

[16] J. Luo, S. Chai, B. Zhang, Y. Xia, J. Gao, G. Zeng, "A novel intrusion detection method based on threshold modification using receiver operating characteristic curve," Concurrency and Computation: Practice and Experience, pp. 5690-5703, 2020.

[17] M. Ahsan, M. Mashuri, M. H. Lee, H. Kuswanto, D.D. Prastyo, "Robust adaptive multivariate hotelling's t2 control chart based on kernel density estimation for intrusion detection system," Expert Systems with Applications, vol. 145, pp. 113105, 2020.

[18] N. Moustafa, J. Slay, G. Creech, "Novel geometric area analysis technique for anomaly detection using trapezoidal area estimation on large-scale networks," IEEE Transactions on Big Data, vol. 5, no. 4, pp. 481–494, 2017.

[19] F. Gottwalt, E. Chang, T. Dillon, "Corrcorr: A feature selection method for multivariate correlation network anomaly detection techniques," Computers& Security, vol. 83, pp. 234–245, 2019.

[20] A. Maroosi, E. Zabbah, H.A. Khabbaz, "Network Intrusion Detection using a Combination of Artificial Neural Networks in a Hierarchical Manner," Journal of Electronical & Cyber Defence, Vol. 8, pp. 89-99, 2020. (In Persian)

[21] R. Jalaei, M.R. Hasani Ahangar, "Detecting Botnets with Timing-Based Covert Command and Control Channels," Journal of Electronical & Cyber Defence, Vol. 7, pp. 1-15, 2019. (In Persian)

[22] C. Jie, L. Jiawei, W. Shulin, Y. Sheng, "Feature selection in machine learning: A new perspective," Neurocomputing, vol. 300, pp. 70-79, 2018.

[23] I. Caturvedi, E. Ragusa, P. Gastaldo, R. Zunino, E. Cambria, "Bayesian network based extreme learning machine for subjectivity detection," Journal of the Franklin Institute, vol. 355, pp. 1780-1797, 2018.

[24] R. Elkan, "Results of the KDD 99 classifier learning," ACM SIGKDD Explorations Newsletter, vol. 1, pp.63-64, 2000.

[25] M. Aldwairi, Y. Khamayseh, M. Al-Masri, "Application of artificial bee colony for intrusion detection systems," Security and Communication Networks, vol. 8, pp. 2730-2740, 2015.

[26] H. Shirazi, Y. Kalaji, "An intelligent intrusion detection system using genetic algorithms and features selection," Majlesi Journal of Electrical Engineering March, vol. 4, pps.33-43, 2010.