معرفی یک روش ترکیبی طبقه بندی-ابتکاری به منظور شناسایی نفوذ و تحلیل رفتارهای مخرب

ساناز شهرکی
دانشجوی مقطع کارشناسی ارشد مهندسی فناوری اطلاعات، دانشگاه سیستان و بلوچستان

فاطمه زهرا شجاعی
دانشجوی مقطع کارشناسی ارشد مهندسی برق گرایش کنترل، دانشگاه شهید باهنر کرمان

چکیده

در این پژوهش یک روش نوین و مقیاس‌پذیر برای تشخیص داده‌های مخرب و نفوذ ناشی از آن ارائه‌شده است. روش معرفی‌شده شامل سه مرحله خرید پیشرفت زمانی، مرور کاربران و مقیاس‌پذیری با کاربرد در حوزه داده‌های کلان است. روش پیشنهادی برای آموزش داده‌ها، زمان به بازه‌های زمانی تقسیم نموده و از اطلاعات مروری[1] کاربران در هر بازه زمانی بهره‌برداری و مورداستفاده قرار می‌گیرد. این روش ساختار ترکیب‌شده شامل نرم‌افزار و سخت‌افزار برای تشخیص داده‌های مخرب و استخراج ویژگی را دربرمی گیرد. برای طبقه‌بندی درروش پیشنهادی از الگوریتم ماشین بردار پشتیبان تغییریافته و برای پیش‌بینی از الگوریتم کلونی باکتری ترکیب‌شده با الگوریتم سیستم ایمنی بدن استفاده‌شده است. نتایج حاصل از این پژوهش نشان می‌دهد روش پیشنهادی در مقایسه با سایر روش‌های مرسوم برای داده‌های کلان بهبود دقت ۹۷.۲ درصد است.

واژگان کلیدی: الگوریتم کلونی باکتری، تشخیص نفوذ، رفتارهای مخرب، کلان داده‌ها.

[1] Log Data

مقدمه

به دلیل تنوع سرویس‌های دیجیتال و رشد فنّاوری هر بخش از سیستم در معرض حمله داده‌های مخرب قرار دارد. با توجه به مقیاس، تنوع و سرعت داده‌های مخرب، نرم‌افزارهای دفاع کننده باید با استفاده از یادگیری ماشین قادر باشند تا حمله‌ها را تشخیص دهند. اولین تشخیص داده‌های مخرب برای تشخیص نفوذ حدود ۴۰ سال پیش توسط دنور انجام پذیرفت [Azmi R, Pishgoo B, Nemati H, 2012]. امروزه تشخیص داده‌های مخرب شبکه سخت‌تر و پیچیده‌تر شده است. ولی مسئله یافتن یک‌راه حل مناسب برای حجم بالای دادهای شبکه هنوز به وجود نیامده است.

تحقیقاتی که در این زمینه انجام‌گرفته است می‌توان به مطالعات موجود در [Azmi R, Pishgoo B, Nemati H, 2012, Rahul, Kedia, P., Sarangi, S., & Monika., 2020, Asrigo K, Litty L, Lie D ,2006, Azmi R, Pishgoo B, Nemati H. 2011] اشاره کرد که برای تشخیص داده‌های مخرب از یادگیری ماشین با زبان جاوا استفاده کرده‌اند و همچنین در تحقیقاتی دیگر می‌توان به تشخیص داده‌های میز کار [Azmi R, Pishgoo B, Nemati H, Azmi R, Marin G, Pishgoo B] داده‌های موبایل [Bello, I., Chiroma, H., Abdullahi, U. A., Gital, A. Y. U., Jauro, F., Khan, A., ... & Abdulhamid, S. I. M. ,2021, Bovet D, Cesati M, Oram A,2002] تشخیص نفوذ شبکه [Kumar, P., Gupta, G. P., & Tripathi, R. ,2021, Rabbani, M., Wang, Y. L., Khoshkangini, R., Jelodar, H., Zhao, R., & Hu, P, 2020, Darvishzadeh N, Azmi R, 2007, Heady R, Luger G, Maccabe A, Servilla S, 1990] تشخیص اسپم [Süzen, A. A, 2021, Kumari, U., & Soni, U., 2007] و تشخیص آدرس‌های جعلی [Ji Z, Dasgupta D ,2004, Ji Z, Dasgupta D ,2004] اشاره نمود.

برخلاف کاربردهای دیگر، یادگیری ماشین که برای تشخیص استفاده می‌شوند مانند تشخیص متن یا چهره که در آن‌ها شکل‌ها و کاراکترهای ثابتی وجود دارد و تشخیص بر اساس آن‌ها انجام می‌شود، داده‌های مخرب الگوی ثابتی ندارد و نیاز به تلاش بیشتری برای شناسایی وجود دارد. درواقع این نوع جستجو باید به‌صورت آنلاین باشد و در هرلحظه به‌روزرسانی انجام دهد تا بتواند الگوهای خود را بسازد. این امر باعث ایجاد یک تأخیر می‌شود و همین تأخیر می‌تواند باعث شود دقت تشخیص داده‌های مخرب پایین بیاید. دستگاه‌های تشخیص نفوذ برای کمک به مدیران امنیتی سیستم در جهت کشف نفوذ و حمله به کار گرفته‌شده است. هدف این دستگاه‌ها تنها جلوگیری از حمله نیست بلکه کشف و شناسایی حملات و تشخیص اشکالات امنیتی در سیستم یا شبکه‌های کامپیوتری و اعلام به مدیر است. دستگاه‌های تشخیص نفوذ در کنار دیواره‌های آتش و به‌صورت مکمل امنیتی مورداستفاده قرار می‌گیرد. برخی از فواید این دستگاه‌ها شامل کارایی بیشتر در تشخیص نفوذ، منبع دانش کاملی از حملات، توانایی رسیدگی به حجم زیادی از اطلاعات، توانایی هشدار نسبتاً بلادرنگ که باعث کاهش خسارت می‌شود.، دادن پاسخ‌های خودکار مانند قطع ارتباط کاربر، افزایش میزان بازدارندگی، توانایی گزارش دهی است.

از فن‌های دیگر می‌توان به معیارهای آماری اشاره نمود. در نوع پارامتریک مشخصات جمع شده بر اساس یک الگوی خاص در نظر گرفته می‌شود و بر اساس مقادیری که باتجربه حاصل‌شده الگو ساخته می‌شود. و مقایسه صورت می‌گیرد در این روش نیز به دلیل پارامتریک بودن بسیاری از داده‌های مخرب را نمی‌تواند شناسایی کند و دقت موردنظر را نمی‌تواند داشته باشد.

از معیارهای دیگر می‌توان به معیارهای آماری غیر پارامتریک اشاره نمود که داده‌های مشاهده‌شده را بر اساس الگوهای استفاده‌شده مشخصی به‌طور قابل‌قبول تعریف می‌کند. اما با الگوهایی که به‌عنوان قانون مشخص‌شده فرق دارد و به‌صورت شمارشی

نیست. متأسفانه در این فن‌ها ایجاد تعداد زیادی هشدار نادرست می‌شود. زیرا الگوهای رفتاری از جانب استفاده‌کنندگان و سیستم بسیار متفاوت است. دستگاه‌های تشخیص مبتنی بر امضا به میان آمدند که قادر به کشف حملات جدید هستند. در این فن‌ها الگوهای نفوذ از پیش‌ساخته شده به‌صورت قانون نگهداری می‌شود به‌طوری‌که هر الگو انواع مختلفی از یک نفوذ خاص را در برگرفته است و در صورت بروز چنین الگویی در سیستم وقوع نفوذ اعلام می‌شود. در این روش‌ها معمولاً تشخیص‌دهنده دارای پایگاه داده‌ای از امضاها یا الگوهای حمله می‌باشند که سعی می‌کنند با بررسی ترافیک شبکه الگوهای مشابه با آنچه را در پایگاه داده خود نگهداری می‌کنند بیابد. این دسته از روش‌ها تنها قادر به تشخیص نفوذهای شناخته‌شده می‌باشند و در صورت بروز حملات جدید در سطح شبکه نمی‌توانند آن‌ها را شناسایی کنند و مدیر شبکه باید همواره الگوی حملات جدید را به سیستم تشخیص نفوذ اضافه نماید.

دستگاه‌های تشخیص نفوذ برنامه‌های نرم‌افزاری هستند که برای تشخیص نفوذ در شبکه‌های هدف طراحی‌شده‌اند. لذا در این تحقیق به بررسی طراحی، اجرا و عملکرد یک تحلیل مقیاس‌پذیر برای تشخیص محتوای مخرب پرداخته و چارچوب ارائه‌شده برای داده‌هایی با مقیاس بالا برای تشخیص داده‌های مخرب و با در نظر گرفتن برگ خرید زمان است . در این طراحی تغییرات برچسب‌ها در هرلحظه چک می‌شود و تمایل به تغییر و تبدیل‌شدن به داده‌های مخرب را مورد ارزیابی قرار می‌دهد. این اندازه‌گیری‌ها برچسب‌های موقتی تأخیر دار را موردتوجه بیشتری قرار می‌دهد. در تحقیقات پیشین هنوز پیاده‌سازی برنامه‌ای که بتواند مدلی از رفتار غیرعادی به‌وسیله روند اجرای کد فراهم کنند و خطای قابل قبولی داشته باشند ارائه نشده است. به‌علاوه این روش‌ها توانایی تشخیص بسیاری از حملات معمول رادارند ولی از تشخیص حملاتی که بر اساس شرایط مسابقه، تخلف در سیاست‌ها و یا جعل هویت هستند عاجز می‌باشند لذا در کارهای پیشین، آشکارسازی داده‌های مخرب، مدیریت کاربران و رفتار داده‌ها به‌صورت جداگانه انجام می‌گرفت که در اینجا همه باهم در یک چارچوب جاسازی‌شده‌اند و می‌تواند کیفیت پژوهش را بالا ببرد.

پژوهش های پیشین

موضوع امنیت درداده های شبکه همواره یکی از موضوعات پیچیده و چالش‌برانگیز بوده است که به دلیل ماهیت آن‌ها باید از ابزارها و روش‌های مناسب و کارآمد برای حفظ امنیت کاربران و اطلاعات آن‌ها استفاده کرد. یکی از مؤثرترین روش‌هایی که در سال‌های اخیر بسیار موردتوجه فعالان امنیت شبکه قرارگرفته است، استفاده از دستگاه‌های تشخیص نفوذ و ارتقای کارایی آن‌ها در مواجهه با حملات مزاحم است. برای داشتن یک IDS کارآمد، با توجه به ماهیت شبکه، یکی از اساسی‌ترین عناصر و ویژگی‌هایی که در تحلیل داده‌های شبکه باید موردتوجه قرار گیرد، داشتن ماهیت داده‌های بزرگ این داده‌ها است. به‌طور دقیق، واضح است که ترافیک اینترنت به‌طورکلی در سال‌های اخیر در جامعه مدرن رشد کرده است و ما انتظار داریم این روند ادامه یابد [Rahul, Kedia, P., Sarangi, S., & Monika. ,2020]. یکی از مهم‌ترین پارامترهایی که هکرها همواره مدنظر داشته‌اند، کسب اطلاعات در مورد وضعیت شبکه‌های کامپیوتری مانند نفوذ به پایگاه‌های داده و شبکه‌های کامپیوتری مورداستفاده در دستگاه‌های دفاعی است. ازاین‌رو، این شبکه‌ها همیشه در معرض حملات خطرناک هستند. از طرفی شبکه‌ها و ماست‌ها در هر ثانیه با حجم زیادی از داده‌ها مواجه می‌شوند. ازاین‌رو، مکانیسم‌های تشخیص نفوذ باید این کوه در حال رشد از داده‌ها را برای الگوهای نفوذی احتمالی از منظر امنیتی استخراج کنند. این محیط و شرایط تشخیص سریع و دقیق نفوذ را دشوار می‌کند. بنابراین، برای شناسایی این‌گونه نفوذها، لازم است یک سیستم تشخیص نفوذ با استفاده از فن‌های داده‌های بزرگ طراحی شود که بتواند این نوع داده‌ها را که ماهیت داده‌های بزرگ دارند در تشخیص دسترسی‌های غیرمجاز به یک شبکه ارتباطی مدیریت کند. بنابراین، در [Asrigo K, Litty L, Lie D ,2006] از

یک روش یادگیری عمیق آگاه از کلان داده برای طراحی یک سیستم تشخیص نفوذ کارآمد و مؤثر برای مقابله با این چالش‌ها استفاده می‌کند. آن‌ها یک معماری خاص از حافظه بلندمدت کوتاه‌مدت (LSTM) را طراحی کردند، و این مدل می‌تواند روابط پیچیده و وابستگی‌های طولانی‌مدت بین بسته‌های ترافیک ورودی را تشخیص دهد. از این طریق می‌توان تعداد هشدارهای کاذب را کاهش داد و دقت سیستم تشخیص نفوذ طراحی‌شده را افزایش داد.

انواع روش‌های شبکه عصبی جهت شناسایی داده‌های مخرب در [Azmi R, Pishgoo B, Nemati H ,2011] استفاده‌شده است. این روش‌ها به نسبت روش‌های پیشین دارای دقت بالاتر و کارایی بهتر است. زیرا شبکه‌های عصبی به‌طور گسترده‌ای به‌عنوان روش مؤثر تطابق طبقه‌بندی الگوها مورداستفاده قرار می‌گیرد. اما حجم محاسبات بالا و سیکل‌های یادگیری طولانی آن‌ها را از خیلی برنامه‌های کاربردی عقب انداخته است. در تشخیص نفوذ شبکه می‌توان برای حملات شناخته‌نشده از شبکه‌های عصبی استفاده کرد. شبکه‌های عصبی مصنوعی برای شناسایی حملات ناشناخته و اجتناب از نفوذهای پنهانی مخرب دارای مزایای بیشتری نسبت به روش‌های پیشین است.

در[Azmi R, Pishgoo B, Nemati H] از انواع الگوریتم‌های بهینه‌سازی جهت تشخیص نفوذ استفاده‌شده است. در این الگوریتم‌ها در مرحله اول اطلاعات اولیه موردنیاز سیستم به همراه قوانینی که برای آن تعریف‌شده است به الگوریتم‌ها داده می‌شود. سپس الگوریتم‌ها در زمانه‌ای بهینه بهترین پاسخ را می‌یابند. از الگوریتم‌های بهینه‌سازی استفاده‌شده می‌توان به الگوریتم ژنتیک، الگوریتم پرندگان و الگوریتم مورچگان اشاره نمود.

محققان در [Azmi R, Marin G, Pishgoo B] جهت تشخیص نفوذ شبکه از ماشین بردار پشتیبان استفاده کرده‌اند. ماشین‌های بردار پشتیبان خصیصه‌های ورودی با مقادیر حقیقی را با نگاشت غیرخطی به فضایی با ابعاد بالاتر می‌برد و با قرار دادن یک مرز خطی داده‌ها را جدا می‌کند. پیدا کردن یک مرز تفکیک برای جداسازی داده‌ها به مسئله بهینه‌سازی درجه دوم تبدیل می‌شود و از مرز خطی برای تقسیم‌بندی استفاده می‌کند.

در مطالعه [Bello, I., Chiroma, H., Abdullahi, U. A., Gital, A. Y. U., Jauro, F., Khan, A., ... & Abdulhamid, S. I. M. ,2021] محققان از درخت تصمیم جهت تشخیص نفوذ استفاده کرده‌اند که در این روش الگوریتمی پیشنهادشده که با ساختن یک درخت تصمیم روی مجموعه‌ای از الگوها یا امضاهای شناخته‌شده از حملات تعداد مقایسه‌های لازم برای شناسایی یک فعالیت مخرب را به نحو چشمگیری کاهش دهد.در این روش تمامی قوانین به‌صورت یک مجموعه درمی‌آیند و به‌عنوان ریشه درخت مطرح می‌شوند.

استفاده از مقوله آنتروپی در تشخیص نفوذ در تحقیق [Bovet D, Cesati M, Oram A. ,2002] انجام‌شده است. به این معنی که میزان خلوص و بی‌نظمی داده‌ها را مورد ارزیابی قرار می‌دهد. در این روش بهره اطلاعات به تشخیص نفوذ موردنظر کمک می‌کند. بهره اطلاعات یک ویژگی مقدار کاهش آنتروپی از طریق ویژگی خاص است. این روش نیز در مقایسه با روش‌های پیشین نتایج مطلوبی داشته است.

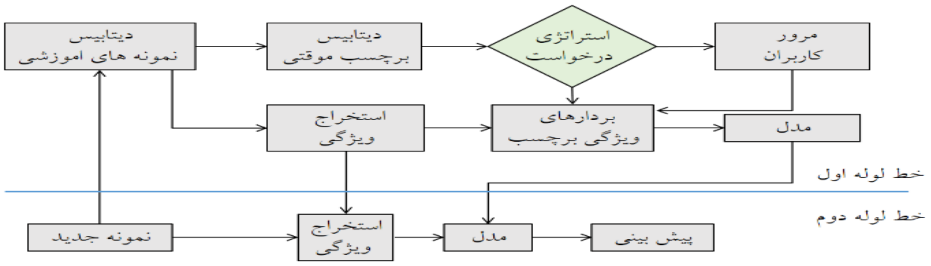
از دیگر تحقیقات انجام‌شده جهت تشخیص نفوذ شبکه استفاده از مدل مارکوف است. در این تحقیقات حملات را به‌عنوان متغیرهای حالت در یک ماتریس حالت/ گذر در نظر می‌گیرند. در این سیستم یک رخداد نابهنجار در نظر گرفته می‌شود هرگاه

احتمال وقوع آ ن رخداد برای حالت قبلی با مقدار وابسته‌اش بسیار کم باشد [Kumar, P., Gupta, G. P., & Tripathi, R,2021].

روش پیشنهادی

در این تحقیق یک روش طراحی چارچوب تجزیه تحلیل مقیاس‌پذیر برای تشخیص داده‌های مخرب ارائه‌شده است. در این طراحی و مدل ارائه‌شده سه برگ خرید پیشرفت زمانی، مرور کاربران و مقیاس‌پذیری لحاظ گردیده است. این طراحی می‌تواند برای داده‌های با حجم بالا مورداستفاده قرار بگیرد. در این روش برای آموزش داده‌ها، زمان به بازه‌های زمانی تقسیم می‌شود و از اطلاعات مروری کاربران در هر بازه زمانی بهره خواهد برد و همچنین داده‌ها آموزش داده می‌شود. برای کار با حجم بالای داده‌ها از فن‌های پیشرفته‌ای استفاده می‌شود و برای مقیاس‌پذیری از فن‌های ذخیره‌سازی برای افزایش سرعت و کاهش حجم محاسبات استفاده‌شده است. این روش یک نوع روش ترکیبی نرم‌افزاری سخت‌افزاری برای تشخیص داده‌های مخرب است. برای دسته‌بندی در این روش پیشنهادی از الگوریتم ماشین بردار پشتیبان تغییریافته و برای عملیات پیش‌بینی از الگوریتم باکتری ترکیب‌شده با الگوریتم سیستم ایمنی بدن استفاده خواهد شد.

در این طراحی تغییرات برچسب‌ها در هرلحظه چک می‌شود و تمایل به تغییر و تبدیل‌شدن به داده‌های مخرب را مورد ارزیابی قرار می‌دهد. این اندازه‌گیری‌ها برچسب‌های موقتی تأخیر دار را موردتوجه بیشتری قرار می‌دهد. نوآوری این تحقیق در طراحی نوین چارچوب پیشنهادی و همچنین روش جدیدی برای استخراج ویژگی از شبکه و روشی جدید برای دسته‌بندی می باشد. این طراحی معماری چارچوب بر مبنای خط لوله انجام‌شده است و دارای دو خط لوله می باشد که پردازش را به‌صورت موازی انجام می‌دهد و همین امر باعث افزایش سرعت روش پیشنهادی است. در این چارچوب یک خط لوله پیش‌بینی تعریف می‌شود که داده‌های مخرب را شناسایی می‌کند و خط لوله آموزشی، مدل بعدی و استخراج ویژگی بعدی را برای کاربران خط لوله پیش‌بینی می‌سازد. در طول هر بازه زمانی استراتژی درخواست همه داده‌های آموزشی را مرور می‌کند و جواب‌هایی را برای مرور کاربران انتخاب می‌کند. نتایج مرور کاربران با داده‌های آموزشی و برچسب‌های موقتی با استفاده از مدل‌های موجود ترکیب می‌شود و داده‌های آموزشی برای مدل بعدی را بسازند. در شکل ۱ معماری چارچوب پیشنهادی را می‌توان مشاهده نمود.



شکل۱- معماری پیشنهادی

همان‌طور که در معماری مطرح‌شده در شکل ۱ مشاهده می‌نمایید استخراج ویژگی و مدل جدید در شروع هر بازه زمانی انجام می‌گیرد. بازه‌هایی که پشت سر هم هستند می‌تواند پیشرفت زمانی را بسازد. بازه‌ها ممکن است در سایز و اندازه باهم فرق داشته باشند . برای مثال ارزیابی ممکن است به‌صورت اتفاقی در اولین بازه ۸۰ درصد نمونه‌ها را به‌عنوان داده‌های آموزشی در نظر بگیرد و ۲۰ درصد نمونه‌ها در بازه‌های بعدی مورد ارزیابی قرار بگیرد. به‌علاوه تقسیم بازه‌ها می‌تواند راه خوبی برای آسان شدن مراحل کار

باشد.

در این مقاله انتخاب ویژگی بر مبنای اطلاعات متقابل است. در تئوری اطلاعات، اطلاعات متقابل MI می‌تواند برای ارزیابی هرگونه وابستگی ، به‌دلخواه بین متغیرهای تصادفی به کار رود. درواقع، MI بین دو متغیر تصادفی X و Y معیاری برای اندازه‌گیری میزان دانش در Y که توسط X عرضه می‌شود است. اگر X و Y مستقل باشند به‌عنوان‌مثال : شامل هیچ اطلاعاتی در مورد Y یا بالعکس نباشند، سپس اطلاعات متقابل آن‌ها صفر است. اطلاعات متقابل دو متغیر تصادفی X و Y مفروض برابر است با:

$$\begin{aligned} I(X:Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X:Y) \end{aligned} \quad (۱)$$

که H(,) آنتروپی است، H(X|Y) و H(Y|X) آنتروپی شرطی است، و H(X:Y) آنتروپی مشترک X وY است که به‌صورت زیر تعریف می‌شود:

$$-H(X) = -\sum_X Px(x) log px(X) \quad (۲)$$

$$H(Y) = -\sum_y Py(y) log\ Py(y) \quad (۳)$$

$$H(X:Y) = -\sum_X \sum_Y Px,y(x,y) log Px,y(x,y) \quad (۳)$$

که $P_{x,y}(x,y)$ تابع چگالی احتمال مشترک و $P_x(x)$ و $P_y(y)$ به ترتیب تابع چگالی حاشیه‌ای xو y است.

$$Py(y) = \sum_y Px,y(x,y) \quad (۵)$$

$$Px(x) = \sum_y Px,y(x,y) \quad (۶)$$

با جایگزین کردن رابطه ۲ و ۴ در معادله ۱ به معادله MI خواهیم رسید.

$$I(X:Y) = \sum_X \sum_Y Px,y(x,y) log \frac{Px,y(x,y)}{Px(x)Py(y)} \quad (۷)$$

ماژول پیش‌بینی خروجی‌ها را از دسته‌بندی کننده ذخیره می‌کند. که این داده‌ها بر روی یک میزبان[2] و در حافظه‌های توزیع‌شده یکسان ذخیره می‌شوند. بدین منظور از الگوریتم باکتری استفاده‌شده است. این الگوریتم ترکیبی از الگوریتم‌های باکتری و الگوریتم سیستم ایمنی بدن است.

الگوریتم بهینه‌سازی تجمعی غذایانی باکتری‌ها یکی از جدیدترین الگوریتم‌های بهینه‌سازی ایده گرفته‌شده از طبیعت است. ایده اصلی در طراحی این الگوریتم استفاده از استراتژی غذایانی باکتری ای کولی در بهینه‌سازی چند توابع با چند بهینه بوده است.

در این بخش از مقاله می بایست مراحل غذایابی صورت می‌گیرد.

غذایابی مساله به دو بخش تقسیم می‌شود.

[2] HOST

ابتدا مهاجم(غذا یابنده) بایستی منبع غذا را یافته و سپس آن‌ها تعقیب کرده و به گروه حمله می‌کند. اهمیت بخش‌های مختلف این روند به رابطه مهاجم و دسته بستگی دارد. بعضی دسته‌ها بزرگ‌اند پس مهاجم نیاز به انرژی بیشتری برای شکاردارند اما در عوض به‌سادگی پیدا می‌شوند

در این تحقیق از روش غذایانی گروهی برای غذایانی استفاده‌شده است. در هنگام غذایانی باکتری، حرکت توسط مجموعه از فلاژل‌ها باقابلیت کشش انجام می‌شود. فلاژل‌ها این امکان را برای
باکتری ای کولی[3] فراهم می‌آورد که چرخیده یا شنا کند. این دو عمل در زمان غذایابی انجام می‌شود. وقتی فالاژل‌ها به سمت عقربه‌های ساعت می‌گردند باعث چرخش سلول می‌شود.

چرخش در محیط‌های سمی (عدم وجود غذا) و یا هنگامی‌که غذا یافت شده است به چشم می‌خورد کمتر است. با چرخش فلاژل‌ها عکس عقربه‌های ساعت، باکتری به اسرعت قابل‌توجهی شنا می‌کند. با توجه به این اعمال باکتری علاقه‌مند به یافتن غذای افزاینده و فرار از محیط‌های سمی است. به‌طورکلی باکتری در محیط‌های دوستانه طولانی‌تر عمل می‌کند.

در هنگامی‌که غذای کافی وجود داشته باشد، طول این باکتری افزایش‌یافته و در دمای مناسب به دو کپی خود تبدیل می‌شود. این عمل باعث به وجود آمدن عمل تولید مجدد در الگوریتم می‌گردد. با توجه به وقوع تغییرات ناگهانی محیطی و یا حمله، پیشرفت شموماتیک ممکن است از بین رفته و یا اینکه گروهی از باکتری‌های به نقطه دیگری منتقل شوند. این اتفاق از بین رفتن و یا پخش شدن در باکتری واقعی اتفاق می‌افتد

به‌صورت خلاصه می‌توان گفت جستجوگرهای ما برای داده‌های مخرب دارای رفتارهای زیر می‌باشند:

رفتار حرکتی باکتری‌ها: که از آن به‌عنوان دوره حیات باکتری‌ها یاد می‌شود. این رفتار شامل Nc تکرار (طول دوره حیات) بوده و در آن باکتری‌ها گام‌هایی برای جست‌وجوی مواد مغزی برمی‌دارند.

اگر درحرکت اول که در جهت ندوم انجام می‌شود تابع هزینه کمتر شده باشد، تا Ns گام دیگر می‌توان در همان جهت جلو رفت به‌شرط آنکه در هر گام کاهش هزینه داشته باشیم. برای حرکت باکتری‌ها از رابطه زیر استفاده می‌شود:

$$\theta^i(j+1,k,l) = \theta^i(j,k,l) + C(i)dlt(i)/(\ dlt(i)^T\ dlt(i))^{0.5} \qquad (۸)$$

که در آن $\theta i(j+1,k,l)$ موقعیت باکتری i در مرحله j+1 از رفتار حرکتی باکتری‌ها و kامین مرحله تولیدمثل و lامین مرحله حذف و پراکندگی است. C(i) گام حرکت و dlt(i)یک بردار رندم D بعدی در بازه [۰،۱] برای تعیین جهت است.

باکتری‌ها در شرایط خاصی ماده‌ای جاذب از خود ترشح می‌کنند که موجب جذب باکتری‌های دیگر به سمت یک ناحیه خاص می‌گردد. بر اساس این ارتباط در رابطه به‌روزرسانی تابع هزینه هر باکتری پس از حرکت آن طبق رابطه زیر باید مقدار $J_{cc}(\theta, \theta i(j,k,l))$ نیز به آن افزوده می‌شود که نماینده‌ای از میزان نیروهای جاذب و دافع بین باکتری‌ها در جمعیت است. (θ یک بردار در فضای D بعدی است.)

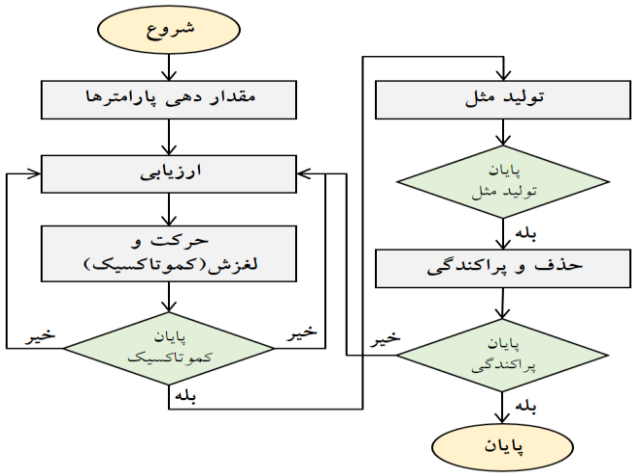
[3] E-Coli

$$J(i,j,k,l) = J(i,j,k,l) + J_{cc}(\theta, \theta^{i}(j,k,l))$$

پس‌ازآنکه دوره حیات باکتری‌ها برای حرکت به پایان رسید میزان سلامت باکتری‌ها که متناسب با میزان مواد مغذی جمع‌آوری‌شده در طول دوره حیات است بر اساس رابطه زیر برای همه باکتری‌ها محاسبه می‌گردد. سپس تعدادی از باکتری‌ها با بیشترین مجموع تابع هزینه می‌میرند (حذف می‌شوند) و به همان تعداد از بهترین باکتری‌ها تکثیر می‌شوند. (به دو باکتری تبدیل می‌شوند.)

$$J^{i}_{health} = \sum J(i,j,k,l)$$

در ازدحام واقعی باکتری‌ها اثر تغییرات محیطی مثل افزایش دما ممکن است خیلی از باکتری‌ها از بین بروند یا به نواحی دیگری بروند. با الهام از این رفتار بعد از تعداد تکرار خاصی از مرحله تولیدمثل هریک از باکتری‌ها حذف‌شده و به مکان دیگری پرتاب می‌گردند (تبدیل به یک باکتری دیگر می‌شود.). فلوچارت این الگوریتم را در شکل ذیل منعکس شده است:



شکل۲- رویکرد باکتری

پس‌ازآنکه عملیات پیش‌بینی داده‌های مخرب را با الگوریتم باکتری انجام دادیم برای افزایش دقت الگوریتم یک‌بار دیگر پیش‌بینی را با الگوریتم سیستم ایمنی بدن انجام می‌دهیم. سیستم ایمنی بدن شامل خصوصیاتی که بالطبع سیستم ایمنی مصنوعی از آن پیروی می‌کند که به‌صورت خلاصه در مورد هرکدام توضیح داده‌شده است:

قابلیت تشخیص الگو توسط آنتی‌بادی‌ها:

این تشخیص الگو با استفاده از یک آستانه انجام می‌شود و زمانی که تحریک الگویی از یک آستانه بالاتر رفت به‌عنوان یک سلول خودی شناخته می‌شود.

تطبیقی بودن سیستم ایمنی با رخدادها و محیط:

بدن انسان با محیط طبیعی خود در ارتباط است و همچنین مواد مفید یا مضر (عوامل بیماری‌زا) که وارد بدن انسان می‌شوند متغیر هستند به همین دلیل سیستم ایمنی به‌صورت پویا با تغییرات برخورد می‌کند و باوجود تغییر عوامل بیماری‌زا را تشخیص داده و با آن‌ها مبارزه می‌کند.

بررسی و ارزیابی کارایی

در این ارزیابی برچسب‌ها را مورد ارزیابی قرار می‌دهیم . به این صورت که ۴ دسته مختلف از داده‌ها را به‌صورت تصادفی انتخاب می‌شود. رخدادهای تکراری نمونه‌ها در ترتیب اسکن اصلی بیانگر چندین اسکن نمونه‌های یکسان در نقاط مختلف در زمان است. در جدول ۱ ارزیابی بین مدیریت برچسب داده‌های آزمودن و آموزش برای انجام عملیات جستجو قابل‌مشاهده است. برچسب‌های نمونه‌های ۲ و ۳ در هر زمان ثابت هستند. و برچسب‌های ۱ و ۴ و۵ با افزایش دانش داده‌ها در حال تغییر است. استفاده از دانش برچسب‌ها می‌تواند کارایی روش را بالاتر ببرد. در اینجا داده‌های ۱ و ۴ بسیار آسان‌تر در بازه زمانی ۳ شناسایی می‌شوند و این نشان می‌دهد که در بازه زمانی ۱و ۲ آموزش‌دیده‌اند. این ارزیابی نشان می‌دهد که دانش برچسب‌ها تا چه اندازه می‌تواند در کیفیت آموزش داده‌ها و نمونه مؤثر واقع شود. برای مثال نمونه ۱ به‌عنوان داده مخرب در اینجا شناسایی‌شده است و در بازه زمانی ۳ به این شناسایی رسیده است. برای ارزیابی برچسب کردن را در پایان هر بازه زمانی انجام دادیم. داده‌های آموزش قبل از داده‌های آزمودن پردازش می‌شوند و برچسب‌های ارزیابی بعد از پایان همه بازه‌ها جمع‌آوری می‌شوند.

جدول۱- بازه های زمانی

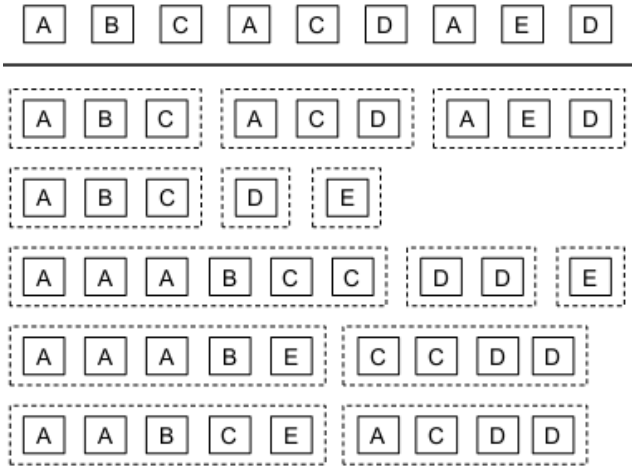
نمونه	بازه زمانی ۱ ۱ ۲ ۳	بازه زمانی۲ ۱ ۳ ۴	بازه زمانی ۳ ۱ ۵ ۴
برچسب در بازه زمانی ۱	− − +		
برچسب در بازه زمانی ۲	− − +	− + −	
برچسب در بازه زمانی ۳	+ − +	+ + −	− − +
برچسب مربوط به ماه در بازه زمانی ۳	+ − +	+ + +	+ + +

در شکل ۳ مقایسه نوع توزیع داده‌ها را با روش‌های بیان‌شده برای مجموعه داده‌های A تا E نشان می‌دهد.

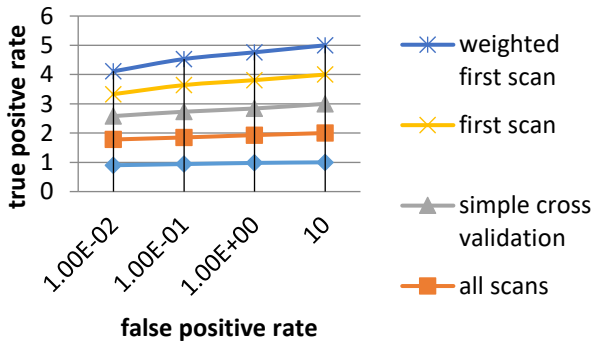
در این شکل خط‌چین‌ها بیانگر بازه‌های زمانی است و خط اول بیانگر ترتیب اسکن اصلی است. و خطاهای بعدی روش‌های اسکن به ترتیب روش all scans، روش first scan only، روش weighted first scan، روش sample cross validationو روش پیشنهادی ما

روش scan cross validation است.

همان‌طور که در شکل ۴ نشان داده‌شده است روش‌های پارتیشن بندی در هر ۵ روش برای تست و آموزش متفاوت است. و مدل پارتیشن بندی برای n بازه زمانی همان مدل برای بازه زمانی n+1 خواهد بود. درروش first scan only هر نمونه دقیقاً در یک بازه زمانی اسکن می‌شود و برای هر نمونه اولین رخداد باقی می‌ماند. در روش weighted first scan اطلاعات نمونه‌های محلی را نگه می‌دارد ولی همه اسکن‌ها در اولین بازه زمانی جای می‌دهد. در روش cross validation داده‌های آموزش و تست با توجه به زمان چیده می‌شوند. این روش نسبت به روش‌های پیشین مناسب‌تر می‌باشد.



شکل۳- پارتیشن بندی



شکل۴ مقایسه روش های پارتیشن بندی

نتیجه گیری

در این تحقیق یک روش برای طراحی چارچوب برای تشخیص داده‌های مخرب با استفاده از الگوریتم‌های یادگیری ماشین سیستم ایمنی مصنوعی و الگوریتم باکتری برای پیش‌بینی داده‌های مخرب و استفاده از الگوریتم DBSCAN برای و طراحی تکنیکی چارچوب استفاده‌شده است. در چارچوب پیشنهادی زمان و برچسب‌ها و مقیاس‌پذیری همچنین مرور کاربران لحاظ گردید و نتایج نشان داد که توضیح قابل قبولی در مقایسه با روش‌های پیشین می‌تواند داشته باشد. این طراحی معماری چارچوب بر مبنای خط لوله انجام‌شده است و دارای دو خط لوله است که پردازش را به‌صورت موازی انجام می‌دهد و همین امر باعث افزایش سرعت روش پیشنهادی است. این برنامه توسط نرم‌افزار آپاچی اسپارک، یک چارچوب محاسباتی و نرم‌افزار متلب انجام‌شده است. همچنین از یک اینترفیس مجازی وب برای مشاهده نتایج استفاده نمویم. این روش را برای آشکارسازی داده‌های مخرب استفاده شده است و دیتاست ما حاوی داده‌های ۳۰ ماه در سایت virus total است که آنالیز استاتیکی و دینامیکی ویروس‌ها را انجام می‌دهد و ما در هرلحظه تغییرات برچسب‌ها را اندازه می‌گرفتیم و به‌محض تبدیل‌شدن به داده‌های مخرب آن‌ها را آشکار ساخته ایم.

منابع

[1] Afzali N, Azmi R, Pishgoo B. A new clonal selection algorithm based on radius regularization of anomaly detectors. Accepted in the 16th CSI international symposium on Artificial intelligence and signal processing. AISP; 2012.

[2] Rahul, Kedia, P., Sarangi, S., & Monika. (2020). Analysis of machine learning models for malware detection. Journal of Discrete Mathematical Sciences and Cryptography, 23(2), 395-407.

[3] Asrigo K, Litty L, Lie D. Using VMM-based sensors to monitor honeypots. In: 2nd international conference on virtual execution environments. VEE; 2006. p. 13e23.

[4] Azmi R, Pishgoo B, Nemati H. Hypervisor-based intrusion detection using artificial immune systems. In: 8th international Iranian ISC conference on information security and cryptology 2011. p. 147e53 [Persian

[5] Azmi R, Pishgoo B, Nemati H. Host based anomaly detection using a combination of artificial immune systems and hypervisor technology. Elsevier Procedia Computer Science, in press-a.

[6] Azmi R, Marin G, Pishgoo B. Operating system formal framework for attack graph creation. Elsevier Procedia Computer Science, in press-b.

[7] Bello, I., Chiroma, H., Abdullahi, U. A., Gital, A. Y. U., Jauro, F., Khan, A., ... & Abdulhamid, S. I. M. (2021). Detecting ransomware attacks using intelligent algorithms: Recent development and next direction from deep learning and big data perspectives. Journal of Ambient Intelligence and Humanized Computing, 12, 8699-8717.

[8] Bovet D, Cesati M, Oram A. Understanding the Linux kernel .Sebastopol, CA, USA: O'Reilly & Associates, Inc.; 2002.

[9] Kumar, P., Gupta, G. P., & Tripathi, R. (2021). Toward design of an intelligent cyber attack detection system using hybrid feature reduced approach for iot networks. Arabian Journal for Science and Engineering, 46, 3749-3778.

[10] Rabbani, M., Wang, Y. L., Khoshkangini, R., Jelodar, H., Zhao, R., & Hu, P. (2020). A hybrid machine learning approach for malicious behaviour detection and recognition in cloud computing. Journal of Network and Computer Applications, 151, 102507.

[11] Darvishzadeh N, Azmi R. Intrusion detection based on system calls analysis. In: 13th international Iranian conference of computer society, Tehran, Iran 2007. p. 200e5[Persian].

[12] Heady R, Luger G, Maccabe A, Servilla S. The architecture of a network level intrusion detection system. Technical report.University of New Mexico, Department of Computer Science 1990.

[13] Süzen, A. A. (2021). Developing a multi-level intrusion detection system using hybrid-DBN. Journal of Ambient Intelligence and Humanized Computing, 12(2), 1913-1923.

[14] Kumari, U., & Soni, U. (2017, October). A review of intrusion detection using anomaly based detection. In 2017 2nd International Conference on Communication and Electronics Systems (ICCES) (pp. 824-826). IEEE.

[15] Ji Z, Dasgupta D. Augmented negative selection algorithm with variable-coverage detectors. In: The IEEE congress on evolutionary computation (CEC'04). IEEE Press; 2004.p. 1081e8.

[16] Ji Z, Dasgupta D. Real-valued negative selection using variable-sized detectors. In: The genetic and evolutionary computation conference (GECCO'04). Berlin/Heidelberg:Springer; 2004. p. 287e98.

Presenting a classification-heuristic hybrid method in order to identify intrusions and malicious behaviors

Sanaz Shahreki

Master's degree student of Information Technology Engineering, University of Sistan and Baluchistan

Fatemehzahra Shojaei

Master's student in electrical engineering, control major, Shahid Bahonar University, Kerman

1-1-

Abstract -۲-۱

In this research, a new and identifiable method for malicious detection and penetration is presented. The introduced method includes three stages of time purchase, user review and identification with application in macro domains. The proposed method for training data is divided into time intervals and exploits and uses the browsing information of users in each time interval. This method includes the combined structure of software and hardware to detect malicious data and features. For classification in the proposed method, the support vector machine algorithm has been modified and for prediction, the bacterial colony algorithm combined with the immune system algorithm has been used. The results of this research show that the proposed method improves the accuracy by 97.2% compared to other conventional methods for big data.

Keywords: Bacterial colony algorithm, intrusion detection, malicious behaviors, big data. . **-۱-۳**