JCWR2020

Cognitive Analysis Based On Machine Learning For Web Logs

Hamid Imani [1], Kourosh Dadashtabar [2]

[1] school of electrical and computer engineering, Malek Ashtar University of Technology, Tehran, Iran
mohandes.imani@gmail.com
[2] school of electrical and computer engineering, Malek Ashtar University of Technology, Tehran, Iran
dadashtabar@mut.ac.ir

Abstract

Logs contain valuable information about user actions on the web that are widely used in a variety of security, industry and science areas. Logs are an excellent resource for determining the health of a system. An enormous amount of logs are generated every day that record users' activities, which is very difficult to analyze in traditional ways, so there is a need for intelligent and cognitive analytics. Machine learning techniques are currently used as an efficient tool in the processing and analysis of web logs. In this paper, we used a 5-layer neural network with 2461 parameters and sigmoid, hyperbolic and relay activation functions, in addition to intelligently detecting cyber threats to optimal results in accuracy and minimizing error function. In analyzing web logs compared to other conventional machine learning methods such as clustering, decision tree, support vector machines, principal component analysis, isolated forest and logistic regression.

Keywords: Log Analysis, Security, Machine Learning, Multilayer Perceptron Neural Network

JCWR2020
کنفرانس بین المللی وب پژوهی
ایران، تهران، ۲۶ خرداد ۱۳۹۹

تحلیل شناختی مبتنی بر یادگیری ماشینی لاگ‌های وب

حمید ایمانی[۱]، کوروش داداش تبار[۲]

[۱] دانشکده برق و کامپیوتر، دانشگاه صنعتی مالک اشتر ،تهران
mohandes.imani@gmail.com

[۲] دانشکده برق و کامپیوتر،دانشگاه صنعتی مالک اشتر ،تهران
dadashtabar@mut.ac.ir

چکیده

لاگ‌ها حاوی اطلاعات ارزشمندی در مورد اقدامات کاربر بر روی وب هستند که کاربردهای فراوانی در حوزه های امنیتی، صنایع و علوم گوناگون دارند. لاگ‌ها یک منبع بسیار عالی برای تعیین سلامت وضعیت سیستم هستند. هر روزه حجم عظیمی از لاگ‌ها که فعالیت‌های کاربران را ثبت می‌کنند تولید می‌شوند که تحلیل آنها با روش‌های سنتی، کاری بسیار دشوار است بنابراین نیاز به استفاده از روش‌های تحلیل هوشمند و شناختی است. در حال حاضر روش‌های یادگیری ماشین به عنوان ابزاری کارآمد در پردازش و تحلیل لاگ‌های وب به کار گرفته می‌شوند. ما در این مقاله با بهره گیری از ساختاری که از شبکه عصبی ۵ لایه با ۲۴۶۱ پارامتر و توابع فعالسازی سیگموید،تانژانت هیپربولیک و رلو استفاده می‌کند، توانستیم علاوه بر هوشمند کردن شناسایی تهدیدات سایبری به نتایج مطلوبی در دقت و کمینه کردن تابع خطا بر روی لاگ های جمع آوری شده از پلتفرم EC2 آمازون با معیار ارزیابی F1 برابر یک، در مقایسه با سایر روش‌های مرسوم یادگیری ماشین از قبیل خوشه بندی،درخت تصمیم،ماشین های بردار پشتیبان،تحلیل مولفه اصلی،جنگل ایزوله و رگرسیون لجستیک دست یابیم.

کلمات کلیدی

تحلیل لاگ، امنیت، یادگیری ماشین، شبکه عصبی چند لایه

۱- مقدمه

با توجه به گسترش روزافزون حملات سایبری، موضوع حفظ و مراقبت از وب سایت ها و پورتال‌ها در محیط اینترنت بسیار مهم است، بطوری‌که هرگونه ایجاد اخلال در آنها توسط مهاجمان می‌تواند خدمت های برخط سازمان را به چالش بکشد. ازاین رو نیاز است بررسی مستمر لاگ های برجا مانده در وب سرورها به منظور پیش بینی، پیشگیری، شناسایی، کشف و خنثی سازی حملات سایبری موردتوجه قرار گیرد. یکی از راهکارهای حفظ امنیت و پایداری وب سایت ها و کاهش خسارات مالی و فنی، تحلیل مستمر ترافیک وب می‌باشد.

ترافیک وب مجموعه درخواست‌ها و پاسخ‌های یک وب است که در وب سرور ثبت و ذخیره می‌گردد. تحلیل ترافیک که تحت عنوان لاگ در وب سرور ثبت می‌شود می‌تواند منجر به شناسایی حملات صورت گرفته به وب سرور باشد. ترافیک وب مجموعه درخواست ها و پاسخ های یک وب است که در وب سرور ثبت و ذخیره می‌گردد.

وب سرور سامانه‌ای است که سایت‌ها بر روی آن قرارگرفته و توانایی پاسخ‌گویی به مرورگر وب و ارسال صفحه درخواستی مرورگر را دارا است. مهاجمان، مسیرها و روش های مختلفی را به کار می‌گیرند تا از طریق وب سایت‌ها و برنامه‌های کاربردی تحت وب، به کشورها، سازمان‌ها و یا کسب وکاری آسیب وارد کنند.

می‌توان با تحلیل لاگ وب، رفتارهای غیر عادی سیستم را به موقع کشف نمود که این موضوع نقش مهمی در مدیریت حادثه سیستم های در مقیاس بزرگ دارد. تشخیص به موقع رفتارهای غیرعادی به توسعه دهندگان سیستم (یا اپراتورها) اجازه می‌دهد تا سریعاً مسائل را مشخص کرده و بلافاصله آنها را برطرف کنند، در نتیجه می‌توانند خرابی سیستم را کاهش دهند[2].

حجم بالای لاگ‌ها ما را از سیستم ها و فناوری‌های مدیریت و تحلیل داده‌های قدیمی به سمت تحلیل هوشمند و شناختی لاگ‌ها سوق می‌دهد. بکارگیری شناخت در تحلیل لاگ‌ها ، شاخه جدیدی به نام تحلیل شناختی لاگ‌ها ایجاد کرده است[4]. تحلیل شناختی لاگ‌ها با تعاملاتی که با محیط پیرامون خود برقرار می‌نماید و با توجه به بازخوردهایی که از محیط می‌گیرد به پیش بینی دقیق تر سلامت شبکه کمک نماید و عملکرد شبکه را بهبود بخشد و به مدیران امنیت شبکه سازمان‌ها در تشخیص هرچه سریعتر خرابی-های احتمالی شبکه کمک نماید. در تحلیل شناختی لاگ وب، سیستم‌های پردازش شناختی می‌بایست بتوانند بطور مستقل و خودمختار مسائل خود را حل کنند[6]. در واقع انتظار داریم این سیستم‌ها بتوانند رویکردی انسانی داشته باشند؛ یعنی در اثر تعامل با محیط اطراف و دنیای درونی خویش، و سعی و خطا قادر باشند نیازهای خود را مرتفع سازند و با استفاده از مکانیزم پیچیده ای به نام یادگیری به تدریج روش‌های هوشمندانه تری را برای حل مسائل خود برگزینند. ارزیابی امنیتی و ایمن بودن شبکه و سیستم‌های مرتبط در مقابل حملات مخرب به یکی از جنبه های حیاتی فعالیت نهادها و سازمان‌ها تبدیل شده است لذا محققین، نهادها و سازمان‌ها به طور فزاینده ای به سراغ راه حل‌های هوشمند و شناختی رفته اند تا با بکارگیری آنها اقدام به جمع آوری و بررسی و تحلیل لاگ‌های داده ، یافتن موارد غیرطبیعی و ایجاد یک رویداد برای یک تحلیل امنیتی نمایند[5]. لاگ یک فایل متنی ساده است که توسط نرم افزاری که روی سرور نصب شده، هرگونه کنش کاربر در تعامل با نظام اطلاعاتی(کلیک کردن با موس یا فشردن کلید بر روی صفحه کلید) را در قالب یک خط داده ثبت می‌کنند. لاگ دربردارنده اطلاعاتی درباره: نام، نشانی آی‌پی ، منطقه زمانی، درخواست دسترسی، نشانی اینترنتی ارجاع شده و یا کدهای خطا است و عموماً در خدمات دهنده وب ثبت است.

تحلیل لاگ وب، فرایندی از انتقال اطلاعات خام لاگ کاربران به اطلاعاتی برای حل مشکلات در مواجه انسان و رایانه است. در سال‌های اخیر، فناوری‌های یادگیری ماشین به سرعت در حال توسعه هستند.

به تازگی، محققان پیشرفت‌های زیادی در تحلیل داده‌های هوشمند با استفاده از شبکه های عصبی ایجاد کرده‌اند[6]. با الهام از این آثار، این مقاله ساختاری را پیشنهاد می‌کند که از شبکه عصبی چند لایه پرسترون برای آموزش لاگ ها استفاده می‌کند. لاگ های وب، ورودی شبکه عصبی پیشنهادی ما در این مقاله هستند. نتایج روش پیشنهادی ما نشان می دهد که محاسبه، تحلیل و پردازش لاگ‌های وب به صورت هوشمندانه با عملکردی بهتر از روش‌های یادگیری ماشین قبلی امکان پذیر است. ما در بخش ۲ این مقاله به مدل ها و روش‌های رایج در یادگیری ماشین پرداخته ایم و سپس در بخش ۳ به کارهای انجام شده در زمینه تحلیل لاگ اشاره نموده ایم و سپس در بخش ۴ روش پیشنهادی خود را بیان نموده و در بخش ۵ نتایج حاصل از پیاده سازی روش‌های مرسوم یادگیری ماشین شامل خوشه بندی،درخت تصمیم،ماشین های بردار پشتیبان،تحلیل مولفه اصلی،جنگل ایزوله و رگرسیون لجستیک و روش پیشنهادی در این مقاله را ارائه نموده‌ایم و در بخش ۶ نتیجه گیری تحقیق را بیان نموده‌ایم.

۲- تحلیل لاگ وب مبتنی بر یادگیری ماشین

در این بخش به روش‌های یادگیری ماشین مرسوم به منظور تحلیل لاگ وب می‌پردازیم.

۲-۱- خوشه بندی

روش خوشه بندی با استفاده از نمونه‌های برچسب دار مدلی جهت توصیف کلاس‌ها ایجاد می‌کند. فرایند دسته بندی در دو مرحله انجام می‌گیرد. در مرحله اول، بر اساس داده‌ها یک مدل دسته بندی ساخته می‌شود. در مرحله دوم چنانچه صحت مدل قابل قبول بود از آن برای خوشه بندی داده‌های جدید استفاده می‌شود.

در واقع این فرایند در دو گام یادگیری و آزمایش انجام می‌پذیرد. گام یادگیری یا آموزش (که در آن یک مدل ساخته می شود) و گام آزمایش (که در آن جهت پیشگویی برچسب‌های کلاس از مدل ساخته شده در گام اول استفاده می‌شود). در گام اول یک خوشه بند با تحلیل مجموعه آموزشی ساخته می‌شود. در ابتدا دقت پیش بینی خوشه بند تخمین زده می‌شود. اگر در گام آزمایش از نمونه‌های آموزشی استفاده کنیم دقت بالایی بدست می‌آوریم چرا که خوشه بند دچار بیش برازش داده‌ها شده است. بنابراین از یک مجموعه آزمایشی باید استفاده کرد که شامل مجموعه نمونه های آموزشی نباشد و در ساخت خوشه بند از آنها استفاده نشده باشد.

۲-۲- درخت تصمیم

درخت تصمیم با ترتیب کردن نمونه‌ها از ریشه به سمت برگ‌های درخت، نمونه ها را دسته بندی می‌کند. در این درخت هر گره نشان دهنده یک ویژگی و هر شاخه مقادیر آن ویژگی را مشخص می‌کند. ویژگی که بیشترین تاثیر در دسته بندی دارد در گره ریشه قرار داده می‌شود. برای دسته بندی هرنمونه ابتدا از ریشه شروع می‌کنیم، به هر ویژگی که می رسیم به سمت شاخه ای از درخت که ویژگی نمونه با آن مطابق است پایین می‌رویم. این فرایند برای زیر درخت ها نیز ادامه می‌یابد تا به دسته بندی کل نمونه برسیم. در هر تکرار گرهای که بیشترین اثر را دارد انتخاب می‌شود. درخت های بدست آمده را نیز می توان به صورت دسته ای از دستورهای اگر- آنگاه نیز نمایش داد تا بررسی آن برای انسان راحت تر گردد.

۲-۳- ماشین های بردار پشتیبان

الگوریتم ماشین بردار پشتیبان با یک نگاشت غیرخطی داده های آموزشی را به یک بعد بالاتر تبدیل می‌کند. این الگوریتم با فرض این که داده‌ها به صورت خطی جدا پذیر باشند ابرصفحه‌ای با بیشترین حاشیه بدست می‌آورد که توسط آن داده های یک کلاس از دیگر کلاس‌ها تفکیک می‌شود.

الگوریتم ماشین بردار پشتیبان این ابرصفحه را با کمک بردارهای پشتیبان که اساساً داده‌های آموزش هستند و حاشیه ها که با کمک بردارهای

پشتیبان تعریف می‌شوند پیدا می‌کند. این الگوریتم در آموزش سرعت پایینی دارد ولی دقت آن بسیار بالا است و کمتر دچار بیش برازش داده‌ها می‌شود.

الگوریتم‌های ماشین بردار پشتیبان خطی را می توان توسعه داد و با ایجاد الگوریتم های ماشین بردار پشتیبان غیرخطی، داده‌هایی که به صورت خطی تفکیک پذیر نیستند را دسته بندی کنیم. ماشین بردار پشتیبان اساساً یک جداکننده دودویی است. یک تشخیص الگوی چند کلاسی می‌تواند به وسیله ترکیب ماشین های بردار پشیبان دوکلاسی حاصل شود.

۲-٤- تحلیل مؤلفه های اصلی

استخراج دانش از حجم زیادی از داده‌ها نیاز به صرف زمان زیادی دارد. روش های کاهش داده ها می‌توانند بدون از دست دادن درستی داده ها و بدون به خطر انداختن نتایج نهایی داده کاوی استفاده شوند. از طرفی کاوش بر روی داده های کمتر هم سریعتر و هم کارآتر است. روش تحلیل مولفه های اصلی بهترین روش برای کاهش ابعاد داده به صورت خطی می‌باشد. یعنی با حذف ضرایب کم اهمیت بدست آمده از این تبدیل، اطلاعات از دست رفته نسبت به روش‌های دیگر کمتر است. در این روش محورهای مختصات جدیدی برای داده‌ها تعریف می‌شود. اولین محور باید در جهتی قرار گیرد که واریانس داده‌ها ماکزیمم شود(یعنی در جهتی که پراکندگی داده‌ها بیشتر است).

دومین محور باید عمود بر محور اول به گونه ای قرار گیرد که واریانس داده‌ها ماکزیمم شود. به همین ترتیب محورهای بعدی عمود بر تمامی محورهای قبلی به گونه ای قرار می‌گیرند که داده ها در آن جهت دارای بیشترین پراکندگی باشند. این روش از مقادیر ویژه و بردارهای ویژه، ماتریس کوواریانس داده‌ها استفاده می‌کند.

۲-۵- شبکه های عصبی

شبکه عصبی مصنوعی روشی عملی برای یادگیری توابع گوناگون نظیر توابع با مقادیر حقیقی، توابع با مقادیر گسسته و توابع با مقادیر برداری هست. انواع شبکه های عصبی برای حل مسائل مختلف یادگیری با نظارت، یادگیری بدون نظارت و یادگیری تقویتی استفاده می‌شوند. شبکه های عصبی بر حسب انواع اتصالات به دو نوع پیشرو و بازگشتی تقسیم می‌شوند.

شبکه‌های عصبی پیشرو با یک لایه مخفی را شبکه عصبی تک لایه پرسپترون و شبکه‌های عصبی پیشرو با بیش از یک لایه مخفی را شبکه عصبی چندلایه پرسپترون می‌نامیم و در آن خروجی نورون‌ها در هر لایه تابعی غیر خطی از خروجی های لایه‌های قبلی است.

برای آموزش (تعیین وزن ها و بایاس‌ها) شبکه های عصبی پیشرو دو راه وجود دارد: روش‌های کلاسیک مانند الگوریتم پس انتشار و روش‌های بهینه سازی هوشمند مانند الگوریتم ژنتیک و الگوریتم بهینه سازی ازدحام ذرات.

شبکه‌های عصبی بازگشتی دارای چرخه‌های جهت دار در ساختار گراف-های ارتباطشان هستند. یعنی با دنبال کردن ارتباطات بین گره‌ها می توان به گره‌ها قبلی و آغازین بازگشت. این گونه از شبکه‌های عصبی بطور خاص برای پردازش داده‌های سری (دنباله ای) مفید هستند. و در آنها هر نورون یا واحد پردازشی قادر به حفظ حالت داخلی یا همان حافظه به منظور حفظ اطلاعات مرتبط با ورودی قبلی می‌باشد. این ویژگی بطور ویژه در کاربردهای مختلف مرتبط با داده‌های سری اهمیت اساسی پیدا می‌کند. ویژگی حفظ حالت درونی یا همان قابلیت حافظه به شبکه کمک می‌کند تا قادر به فهم و کشف ارتباط بین لغات مختلف در دنباله‌های طولانی تر باشد. ایده اصلی پشت این نوع از معماری، بهره برداری از این ساختار سری داده است.

۲-٦- جنگل ایزوله

ایده‌ی اصلی در الگوریتم‌هایی که با نام جنگل شناخته می‌شوند این است که این الگوریتم‌ها از تعدادی درخت (مانند درخت تصمیم) استفاده می‌کنند و با کمک این درختان به یک جمع‌بندی کلی می‌رسند. در الگوریتم جنگل تصادفی، ایده‌ی کلی، به این صورت است که می‌تواند دسته بندی را بر روی مجموعه‌ی داده انجام دهد. اما ایده‌ی الگوریتم جنگل ایزوله کمی فرق می‌کند. این الگوریتم که برای به دست آوردن داده‌های پرت به وجود آمده است می‌تواند داده‌هایی را که از داده‌های دیگر جدا (و تنها) هستند شناسایی کرده و به عنوان داده‌های پرت علامت بزند. الگوریتم جنگل ایزوله یک ویژگی (یک بُعد) را به صورت تصادفی انتخاب می‌کند و سپس یک مقدار تصادفی بین کمینه و بیشینه‌ی آن انتخاب کرده و با یک خط جداساز آن بُعد را جدا می‌کند. این جداساز به صورت تصادفی، ویژگی را انتخاب کرده و با یک خط تصادفی این ویژگی را به دو قسمت تبدیل می‌کند. با این کار می‌توان یک درخت دودویی ساخت که برای جداکردن داده‌ها کار می‌کند. این درخت در ریشه، داده‌ها را به دو قسمت تقسیم می‌کند. هر بار، یک ویژگی به صورت تصادفی انتخاب می شود و آن ویژگی با یک خط جداکننده، به قسمت‌های مختلفی تقسیم می کند تا درخت نظیر آن تولید شود.حال دوباره الگوریتم، یک ویژگی (بُعد) تصادفی را انتخاب می‌کند. و دوباره خطی برای جداسازی آن ویژگی به صورت تصادفی می‌کشد. در واقع ما آنقدر این تقسیم‌بندی را انجام دادیم تا بلاخره یک نقطه، تنهای تنها، در یکی از محوطه‌ها پیدا شود.

پس برای به دست آوردن نقاط تنها بایستی درخت را خیلی بیشتر از این ادامه دهیم. در این‌جاست که به ایده‌ی اصلی جنگل ایزوله می‌رسیم.

۲-۷- رگرسیون لجستیک

همان طور که می‌دانیم در رگرسیون خطی، متغیر وابسته یک متغیر کمی در سطح فاصله‌ای یا نسبی است و پیش بینی کننده ها از نوع متغیرهای پیوسته، گسسته یا ترکیبی از این دو هستند. اما هنگامی که متغیر وابسته در کمی نباشد، یعنی به صورت دو یا چند مقوله‌ای باشد، از رگرسیون لجستیک استفاده می‌کنیم که امکان پیش‌بینی عضویت گروهی را فراهم می‌کند. در تحلیل رگرسیون لجستیک، همیشه یک متغیر وابسته و معمولا مجموعه ای از متغیرهای مستقل وجود دارند. با این که رگرسیون لجستیک در مقایسه با رگرسیون خطی پیش فرض های کمتری دارد.

۳- کارهای مرتبط

نسار و همکارانش پی برده‌اند که پیش بینی خرابی های سیستم بر اساس لاگ‌ها امکان پذیر است. او و همکارانش شش روش تشخیص ناهنجاری برمبنای لاگ را مرور و ارزیابی کرده اند[3]. بنابراین، اگر ما بتوانیم مطالعه بیشتری در زمینه لاگ که از الگوریتم هایی مانند جمع آوری داده، داده کاوی و یادگیری ماشین استفاده می کند انجام بدهیم، می توانیم به یک تحلیل وابستگی معناشناختی سطح بالاتری برسیم و حتی رویدادهای آینده را پیش بینی کنیم. مولار در سال ۲۰۱۲ موضوع کشف ادله جرم در برنامه‌های کاربردی وب را با استفاده از روش‌های آماری و تکنیک‌های آموزشی ماشینی مبتنی بر مدل‌های پنهان مارکوف پیشنهاد داده است. زی و همکاران در سال ۲۰۱۶ مدل پنهان نیمه مارکوف را برای ثبت رفتار کاربران وب اعمال کردند. شی جین و همکاران، سامانه تشخیص نفوذ جدید بر پایه طبقه بندی سلسله مراتبی و ماشین‌های بردار حمایتی را در سال ۲۰۱۱ مورد بررسی و پژوهش قراردادند. در این مطالعه، سامانه تشخیص نفوذی بر پایه ماشین بردار پشتیبان ارائه شده است که الگوریتم طبقه بندی سلسله مراتبی، روند انتخاب مشخصه ساده و تکنیک ماشین بردار پشتیبان را با یکدیگر ترکیب می کند[۱].

برای بررسی حملات وب می‌توان از لاگ‌های شبکه، لاگ‌های وب سرور استفاده کرد. گروهی از محققین دانشگاه سانتیاگو به این نتیجه رسیدند که تعدادی از الگوریتم‌های دسته بندی یعنی جنگل تصادفی و ماشین‌های بردار پشتیبان و شبکه های عصبی مصنوعی بهترین عملکرد رادارند و به جز در موارد خاص، می‌توانیم در اکثر مسائل دنیای واقعی از همین چند الگوریتم اصلی و مناسب استفاده کنیم و به نوعی این الگوریتم‌ها رامی توان الگوریتم-های برتر نامید[۱].

بیپلاب دبناچ و همکارانش یک سیستم تحلیل لاگ تحت عنوان لاگ لنز را ارائه نمودند که در آن ناهنجاری‌ها بطور خودکار و با حداقل تعاملات انسانی، از طریق لاگ‌ها و توسط روش‌های یادگیری ماشین بدون نظارت به صورت زمان حقیقی تشخیص داده می‌شوند[13].

ندا افضلی‌سرشت و همکارانش در مقاله خود روشی را برای تحلیل لاگ‌های وب با استفاده از علوم شناختی و روش‌های یادگیری ماشین برای پایش تهدیدات سایبری در مرکز عملیات امنیت ارائه نمودند که در مقایسه با گزارش کارشناسان انسانی مرکز یاد شده از دقت بالایی برخوردار بود [14].

٤- روش پیشنهادی

مراحلی که در تهیه و تحلیل لاگ به کار گرفته می‌شوند عبارتند از جمع آوری ، آماده سازی و تجزیه و تحلیل لاگ‌ها.

ما در این مقاله به منظور تحلیل لاگ، ابتدا لاگ‌ها را از پلتفرم EC2 آمازون جمع آوری کردیم سپس لاگ‌ها را که یک فایل متنی می باشد تجزیه نموده و سپس لاگ‌های تجزیه شده در مرحله قبل را به بردارهای ویژگی عددی برای ورود به روش‌های یادگیری ماشین تبدیل کردیم و در آخر به تحلیل لاگ پرداختیم.

با مطالعه و بررسی مقالات و تحقیقات [1-14] ، روش پیشنهادی در این مقاله استفاده از شبکه عصبی چند لایه پرسپترون به منظور پردازش و تحلیل لاگ‌های کلان داده در تحلیل شناختی لاگ‌های وب می‌باشد.

در این قسمت به تشریح نتایج حاصل از اجرای روش پیشنهادی مبتنی بر شبکه عصبی چند لایه پرسپترون بر روی لاگ‌های وب کسب شده بر بستر هدوپ پرداخته شده است. پیاده سازی‌های گوناگونی با استفاده از شش روش مبتنی بر یادگیری ماشین شامل تحلیل مولفه اصلی، ماشین بردار پشتیبان، رگرسیون لجستیک، جنگل ایزوله، درخت تصمیم، خوشه بندی به منظور ارزیابی پارامترهای مختلف بر روی لاگ‌های کلان داده بستر هدوپ صورت گرفته است که نتایج بدست آمده از آن‌ها در زیر به تفصیل بیان شده است.

روش پیشنهادی ما در این مقاله، یک شبکه عصبی چند لایه پرسپترون است. مسئله اصلی در طراحی این شبکه تعیین تعداد مناسب لایه های پنهان و تعداد نورون‌های پنهان در لایه‌های میانی است. بنابر روش آزمون و خطا تعداد لایه میانی ۳ لایه در نظر گرفته شده است.پارامتر دیگر، تعداد نورون‌ها درهر لایه است. نورون‌های لایه‌های میانی در شبکه به عنوان تشخیص دهنده الگو عمل می‌کنند. بنابراین تعداد نورون‌ها در لایه پنهان نقش عمده‌ای در قدرت شبکه دارد. کم بودن تعداد نورون‌ها قدرت تجزیه وتحلیل و به دنبال آن دقت عددی خروجی را کاهش می‌دهد. و شبکه نمی‌تواند نگاشت غیرخطی بین ورودی و خروجی را با دقت لازم منعکس کند. از سوی دیگر زیاد بودن بیش ازحد تعداد نورون‌های لایه‌های میانی منجر به تولید نگاشتی غیرخطی و پیچیده شده که در این حالت سیستم داده‌های آموزشی را به جای تجزیه و تحلیل به خوبی یاد می-گیرد، اما در مقابل داده‌های جدید عملکرد مناسبی ندارد و در واقع قدرت تعمیم خود را از دست می‌دهد. برای غلبه بر این مشکل باید تعداد نورون‌ها به گونه‌ای انتخاب شوند که شبکه قدرت کافی و نه بیش از حد برای تولید نگاشت بین ورودی و خروجی داشته باشد. در این مقاله تعداد مناسب نورون-های لایه‌های میانی برای دستیابی به بهترین پیش بینی ها و کمترین میزان خطا در شبکه بر مبنای روش آزمون و خطا برابر ۲۴۶۱ و با پنج تابع فعالسازی به صورت جدول (۱) می باشد.

جدول (۱) : مشخصات شبکه عصبی پیشنهادی

شماره لایه	تعداد نورون	تعداد پارامتر	تابع فعالسازی
۱	۱۰	۱۵۰	سیگموید
۲	۲۰	۲۲۰	سیگموید
۳	۵۰	۱۰۵۰	تانژانت هیپربولیک
۴	۲۰	۱۰۲۰	رلو
۵	۱	۲۱	سیگموید
جمع کل	.....	۲۴۶۱	.....

۵- نتایج

ما شش روش یادگیری ماشین شامل تحلیل مولفه‌های اصلی،ماشین بردار پشتیبان، رگرسیون لجستیک، جنگل ایزوله، درخت تصمیم و روش پیشنهادی در این مقاله را در محیط برنامه نویسی پایتون و در ماشینی با مشخصات پردازنده corei-7 , و گرافیک Nvidia GTX1080 GPU پیاده سازی نمودیم، و به منظور تبدیل لاگ‌های خام به رویداد لاگ از لاگ پارسر برخط متن باز منتشر شده توسط آقای شیلین هی و همکارانش استفاده نمودیم[4].

با عنایت به اینکه در بیشتر مقالات مطالعه شده از سه پارامتر Precision، Recall و F1 جهت ارزیابی روش های یادگیری ماشین استفاده شده است لذا ما نیز در این مقاله از این سه معیار جهت ارزیابی روش‌های یادگیری ماشین شامل تحلیل مولفه‌های اصلی،ماشین بردار پشتیبان، رگرسیون لجستیک، جنگل ایزوله، درخت تصمیم و روش پیشنهادی همان گون که در جدول(۲) قابل مشاهده است استفاده کرده‌ایم.

تابع ضرر، معیاری برای سنجش مناسب بودن مدل پیشنهادی از نظر قابلیت و توانایی در پیشگویی مقدارهای جدید است که در شکل (۱) قابل ملاحظه است.

ما در این مقاله به بررسی برخی از روش‌های رایج یادگیری ماشین و مقایسه آن‌ها با مدل پیشنهادی خود پرداختیم لذا همان‌گونه که در شکل (۲) ملاحظه می گردد الگوریتم مدل پیشنهادی در این مقاله از دقت خوبی برخوردار می‌باشد.

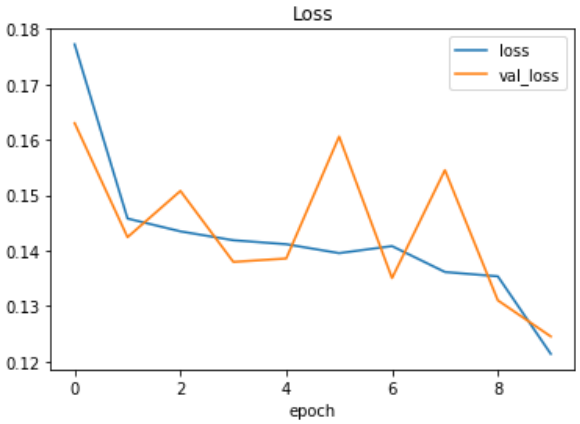
٦- نتیجه گیری

در این مقاله، مشخصات یک شبکه عصبی چند لایه به منظور تحلیل شناختی لاگ‌های وب بیان شد. مهمترین مشخصات عبارتند از: تعدادلایه ، تعداد نورون و توابع فعالسازی.

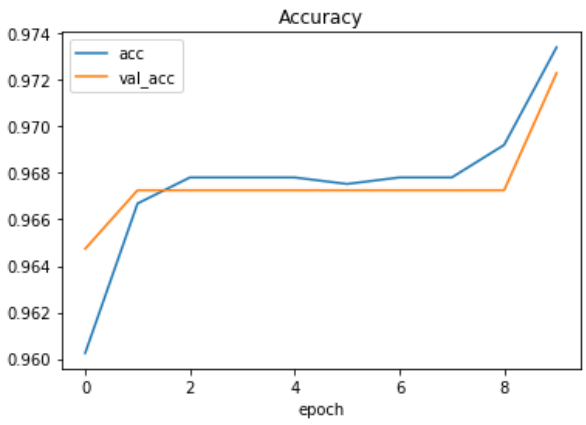
روش‌های یادگیری ماشین روش های تجربی هستند و برای یافتن بهترین نتیجه ممکن باید روش های مختلف با پارامترهای مختلف مورد آزمایش قرار گیرند. با توجه به آزمایشات انجام شده روی مجموعه داده مورد استفاده در این پژوهش، می‌توان مزیت روش پیشنهادی را، مصالحه در تعداد پارامترهای شبکه عصبی و دقت و سرعت در مقایسه با سایر روش‌های یادگیری ماشین مورد استفاده در این مقاله دانست که با توجه به معیار ارزیابی F1برابر یک، روش پیشنهادی از دقت بالاتری در شناسایی تهدیدات سایبری برخوردار است. از آنجایی که روش پیشنهادی در این مقاله دارای چرخه‌های جهت دار در ساختار خود برای دنبال کردن ارتباطات بین گره‌ها قبلی و آغازین و همچنین حافظه به منظور حفظ اطلاعات مرتبط با ورودی قبلی و بالا بردن جنبه شناختی نمی‌باشد لذا می توان از شبکه های عصبی بازگشتی با حافظه طولانی کوتاه مدت که دارای حافظه و چرخه‌های جهت دار در ساختار خود می‌باشد برای کارهای آینده استفاده نمود.

جدول (۲) : مقایسه روش‌های یادگیری ماشین با روش پیشنهادی

مدل	معیار ارزیابی		
	Precision	Recall	F1
تحلیل مولفه های اصلی	0.966	0.363	0.528
ماشین بردار پشتیبان	0.986	0.433	0.602
رگرسیون لجستیک	0.989	0.365	0.535
جنگل ایزوله	0.986	0.433	0.602
درخت تصمیم	0.985	0.427	0.596
خوشه بندی	0.967	0.561	0.710
روش پیشنهادی	1.000	0.433	0.601



شکل (۱) : نمودار تابع ضرر حاصل از داده‌های آزمایش بر روی شبکه عصبی پیشنهادی



شکل (۲) : نمودار دقت حاصل از داده‌های آزمایش بر روی شبکه عصبی پیشنهادی

مراجع

[۱] یادگاری، وحید، **شناسایی جرائم سایبری از طریق همبسته سازی رخدادهای وب** ، کارشناسی ارشد، تربیت مدرس، تهران، ۲۳-۲۱، تابستان ۱۳۹۵.

[2] Mauro Coccoli, Paolo Maresca, Lidia Stanganelli, The role of big data andcognitive computing in the learning process, J. Visual Lang. Comput. 38 ,2017, pp. 97–103.

[3] Fares A. Nassar, Dorothy M. Andrews, A methodology for analysis of failure prediction data, in: IEEE Real-Time Systems Symposium, 1985, pp. 160–166

[4] Shilin He, Jieming Zhu, Pinjia He, Michael R. Lyu, Experience report: Systemlog analysis for anomaly detection, in: IEEE International Symposium onSoftware Reliability Engineering, 2016, pp. 207–218.

[5] Adam Oliner, Archana Ganapathi, Wei Xu, Advances and challenges in loganalysis, Commun. ACM 55 (2) (2012) 55–61.

[6] Yen, Steven, and Melody Moh. "Intelligent log analysis using machine and deep learning." Machine Learning and Cognitive Science Applications in Cyber Security. IGI Global, 2019. 154-189.

[7] Renta Ivلncsy, Istvلn Vajk, Frequent pattern mining in web log data, Acta Polytech. Hung. 3 (1) (2006) 223–228.

[8] Robiah Yusof, Siti Rahayu Selamat, Shahrin Sahib, Intrusion alert correlation technique analysis for heterogeneous log, Int. J. Comput. Sci. Netw. Secur. 59 (9) , 2008, pp. 132–138.

[9] H. Asifiqbal, Nur Izura Udzir, Ramlan Mahmod, Abdul Azim Abd. Ghani, Filtering events using clustering in heterogeneous security logs, Inf. Technol. J. 10 (4) , 2011.

[10] Pinjia He, Jieming Zhu, Shilin He, Jian Li, Michael R. Lyu, Towards automatedlog parsing for large-scale log data analysis, IEEE Trans. Dependable SecureComput. PP (99) ,2017, pp. 1–1.

[11] Hemant Hingave, Rasika Ingle, An approach for MapReduce based log analysis using Hadoop, in: International Conference on Electronics and Communication Systems, 2015, pp. 1264–1268.

[12] Sandeep Kumar Dewangan, Shikha Pandey, Toran Verma, A distributed framework for event log analysis using MapReduce, in: International Conference on Advanced Communication Control and Computing Technologies, 2017, pp. 503–506.

[13] B. Debnath et al., "LogLens: A Real-Time Log Analysis System," 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS), Vienna, 2018, pp. 1052-1062.

[14] N. Afzaliseresht, Y. Miao, S. Michalska, Q. Liu and H. Wang, "From logs to Stories: Human-Centred Data Mining for Cyber Threat Intelligence," in IEEE Access, vol. 8, pp. 19089-19099, 2020.