Scientific Journal of Electronical & Cyber Defence
Vol. 10, No. 4, Winter 2023, Serial No. 40

An Improved Method for Malware Attack Detection in Cloud Computing Using Collective Learning

M. Hesabi, M. Deypir*

* Associate Professor, Shahid Sattari Aviation University, Tehran, Iran

(Received: 21/12/2021, Accepted: 29/01/2022)

ABSTRACT

Nowadays, detecting unusual events in the network has been the subject of many researches. Network traffic is huge and very large, and this leads to high data size and increased noise, which makes it very difficult to extract meaningful information to detect abnormal events. Early detection of attacks improves the stability of a system. Each attack is a type of specific behavior; But some attacks may behave similarly and differ only in some features. This article presents a new way to detect malware and attacks in the cloud computing environment. In this method, data clustering separates the data from each other to provide better conditions for model construction by balancing the data in different classes. This research uses a combination of Adabost, Random Forest and Bosted Gradient Tree algorithms as collective learning to improve malware detection in cloud computing. In order to combine boosted learners and build a higher level model, the voting mechanism is used. In the proposed model, collective learning, using the strengths of various algorithms, creates a useful, high-performance system for detecting malware in cloud computing. By applying the proposed method on real data, it was observed that the accuracy of the proposed method is equal to 99.96%, its accuracy is equal to 99.97% and its recall is equal to 99.95% which compared to previous methods, it has a noticeable advantage, but its computational complexity has not changed much.

Keywords: Malware Attacks, Cloud Computing, Classification, Clustering, Collective Learning.

* Corresponding Author Email: Mdeypir@gmail.com

نشریه علمی "پدافند الکترونیکی و سایبری"
سال دهم، شماره ۴، زمستان ۱۴۰۱، ص ۳۹- ۳۳

علمی – پژوهشی

یک روش بهبودیافته جهت تشخیص حملات بدافزار در رایانش ابری با استفاده از یادگیری جمعی

محسن حسابی[۱]، محمود دی پیر[۲*]

۱- کارشناسی ارشد، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، ۲- دانشیار، دانشگاه هوایی شهید ستاری، تهران، ایران

(دریافت: ۱۴۰۰/۰۹/۳۰، پذیرش: ۱۴۰۰/۱۱/۰۹)

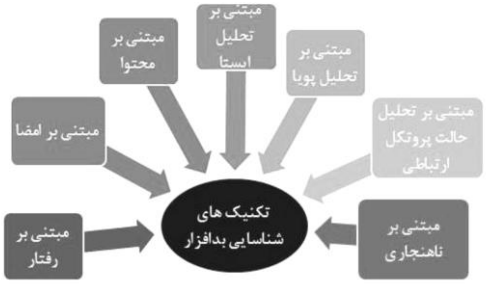
چکیده

امروزه تشخیص اتفاقات غیرعادی در شبکه، موضوع بسیاری از پژوهش‌ها قرار گرفته است. ترافیک شبکه گسترده و بسیار حجیم است و این مسئله منجر به ابعاد بالای داده و افزایش نویز شده و سبب می‌شود که استخراج اطلاعات معنادار برای تشخیص اتفاقات غیرعادی بسیار مشکل گردد. تشخیص به‌موقع حملات، پایداری یک سیستم را بهبود می‌بخشد. هرکدام از حملات گونه‌ای از یک رفتار خاص است؛ اما برخی از حملات ممکن است رفتاری مشابه داشته و فقط در پاره‌ای از ویژگی‌ها متفاوت باشند. در این مقاله، روشی نوین به‌منظور تشخیص بدافزارها و حملات در محیط رایانش ابری ارائه شده است. در این روش، خوشه‌بندی داده‌ها، داده‌ها را از یکدیگر تفکیک می‌نماید تا با متوازن‌سازی داده‌ها در کلاس‌های مختلف، شرایط بهتری برای ساخت مدل فراهم گردد. این پژوهش از ترکیب الگوریتم‌های آدابوست، جنگل تصادفی و درخت گرادینت بوستد به‌صورت یادگیری جمعی به‌منظور بهبود تشخیص بدافزارها در رایانش ابری استفاده می‌کند. به‌منظور ترکیب یادگیرنده‌های جمعی و ساخت یک مدل سطح بالاتر، از مکانیزم رأی‌گیری استفاده می‌گردد. در مدل پیشنهادی، یادگیری جمعی با استفاده از نقاط قوت الگوریتم‌های مختلف، یک سیستم مفید با عملکرد بالا را برای شناسایی بدافزار در رایانش ابری ایجاد می‌نماید. با شبیه‌سازی روش پیشنهادی روی داده‌های واقعی مشاهده گردید که میزان دقت روش پیشنهادی برابر با ۹۹/۹۶٪، میزان صحت آن برابر با ۹۹/۹۷٪ و میزان فراخوانی آن برابر با ۹۹/۹۵٪ است که نسبت به روش‌های گذشته برتری محسوسی دارد و در عین حال پیچیدگی محاسباتی آن نیز تغییر چندانی نداشته است.

کلیدواژه‌ها: حملات بدافزار، رایانش ابری، دسته‌بندی، خوشه‌بندی، یادگیری جمعی

۱- مقدمه

در عصر حاضر با بستر و محیط جدیدی به نام رایانش ابری روبرو هستیم که در این محیط سرویس‌ها از طریق اینترنت به اشتراک گذاشته می‌شوند. اگرچه اشتراک منابع یکی از فواید اساسی رایانش ابری است، همین موضوع می‌تواند زمینه‌ساز یک مخاطره امنیتی بزرگ برای سازمان شود، زیرا فعالیت‌های غیرقانونی که ممکن است توسط دیگران با استفاده از منابع مشترک انجام شود، مستقیماً به شهرت و اعتبار سازمان آسیب وارد می‌کند [۱]. مخاطره شدیدتر آن است که یکی از سازمان‌های مشتری خدمات ابر، به دلیل رخداد یک حمله یا در دیگر شرایط غیرعادی، به دارایی‌های دیگر مشتریان دسترسی پیدا کند [۳-۲]. بدافزارها یکی از انواع حملاتی هستند که در رایانش ابری می‌توانند اجرا شده و به این فناوری تزریق می‌گردند. وجود بدافزارها در محیط رایانش ابری موجب زیان‌های بسیاری شده است. برای شناسایی بدافزارها دیدگاه‌های مختلفی وجود دارد که در شکل (۱) خلاصه شده‌اند.



شکل (۱). دسته‌بندی روش‌های شناسایی بدافزارها

تمرکز این مقاله بر تشخیص بدافزار مبتنی بر ناهنجاری است. مزیت عمده روش‌های مبتنی بر تشخیص ناهنجاری این است که می‌تواند با صرف کمترین هزینه، انواع مختلف و ناشناخته‌ای از حملات را که الگوی آن‌ها قبلاً مشاهده نشده را تشخیص دهد. با وجود روش‌های متعددی که در این زمینه مطرح گردیده، اما هنوز این روش‌ها نتوانسته‌اند با دقت قابل قبولی این حملات را تشخیص دهد. ازاین‌رو در این تحقیق به کمک روش‌های یادگیری جمعی که برای تشخیص ناهنجاری به کار می‌رود اقدام به کشف این نوع حملات و بدافزارها نموده و تلاش می‌کنیم از نفوذ آن‌ها جلوگیری نماییم. در این مقاله روش جدیدی بر مبنای

* رایانامه نویسنده مسئول: Mdeypir@gmail.com

یادگیری جمعی و رأی‌گیری از دسته‌بندهای جمعی، ارائه شده است. آزمایش‌های انجام‌شده نشان‌دهنده برتری روش پیشنهادی نسبت به دسته‌بندهای جمعی ارائه شده در تحقیقات گذشته است. در بخش بعد به ادبیات و پیشینه تحقیق پرداخته شده است. روش پیشنهادی را در بخش ۳ معرفی می‌شود. در بخش ۴ روش پیشنهادی نسبت به روش‌های مشابه گذشته ارزیابی و مقایسه شده است. این مقاله در بخش ۵ نتیجه‌گیری شده است و در آن بخش به برخی از کارهای آینده، اشاره شده است.

۲- ادبیات و پیشینه تحقیق

واژه بدافزار معادل malware انگلیسی است که یک خلاصه برای Malicious Software یا نرم‌افزار بدخواه می‌باشد. واژه بدافزار به ویروس، کرم، تروجان و هر برنامه دیگری که با نیت اعمال خرابکارانه ایجاد شود، اطلاق می‌شود. فن‌آوری‌های مجازی‌سازی، به‌ویژه ماشین‌های مجازی (VM) به‌طور گسترده مورداستفاده قرار می‌گیرند و در هسته محاسبات ابری قرار دارند. درحالی‌که سیستم‌عامل‌های مختلف می‌توانند در بالای نمونه‌های VM اجرا شوند، در محیط‌های ابری عمومی، سیستم‌عامل لینوکس در اغلب مواقع استفاده می‌شود. به دلیل رواج آن‌ها، سرورهای مجازی مبتنی بر لینوکس به هدفی جذاب برای حملات سایبری تبدیل شده‌اند که این حملات عمدتاً توسط بدافزارهای پیچیده طراحی می‌شوند. هدف این حملات ایجاد آسیب، خرابکاری در عملیات، به دست آوردن داده یا کسب سود مالی است. بدافزارها استراتژی‌های مختلفی برای انجام عملیات خرابکارانه خود استفاده می‌کنند که در این بخش توصیف مختصری از آن‌ها بیان می‌کنیم. اولین استراتژی که معرفی می‌کنیم، مبهم کردن کد است. در این روش توسعه‌دهندگان بدافزار فعالیت‌هایی را بر اساس روش‌های مبتنی بر امضا جهت جلوگیری از شناسایی بدافزار بیان می‌کنند. این فعالیت‌ها شامل اضافه کردن فرامین اشتباه و پرش‌های غیرلازم و غیره می‌باشند [۴]. دومین استراتژی، رمزنگاری کد است. این‌گونه برنامه‌های مخرب از مکانیسم دفاعی رمزنگاری برای خودشان یا فعالیت‌های مخربشان استفاده می‌کنند. برنامه‌های مخرب رمزنگاری‌شده مجموعه‌ای پیچیده از الگوریتم‌های رمزگشا و رمزنگاری و کلیدهای رمزنگاری و کد مخرب رمزنگاری شده است [۴]. وقتی‌که برنامه مخربی اجرا می‌شود الگوریتم راهنما و رمزگشا برای رمزگشایی از بخش مخربش استفاده می‌کند. استراتژی بعدی الیگومورفی است. برنامه‌های مخربی که از این استراتژی برای رمزنگاری به‌عنوان مکانیسم دفاعی برای محافظت از خودشان استفاده می‌کنند، قادر هستند الگوریتم رمزنگاری‌شان را در مدت محدودی تغییر دهند. به‌عنوان مثال ویروسی که حلقه‌هایی به تعداد محدود و کوچک و رمز گشاینده مختلفی دارد. استراتژی پلی مورفی چهارمین مورد است. برنامه‌های مخربی که از این استراتژی استفاده می‌کنند، معمولاً خودشان را با یک الگوریتم رمزنگاری، رمزنگاری می‌کنند؛ بنابراین با هر آلودگی کلیدی با رمز نگارنده‌های مختلف مورداستفاده قرار می‌گیرد. همچنین می‌توانند تعداد نامحدودی الگوریتم‌های رمزنگار را برای جلوگیری از شناسایی شدن استفاده کنند [۵]. استراتژی متامورفی به‌عنوان پنجمین استراتژی در اینجا معرفی می‌گردد. برنامه‌های مخربی که از این نوع استراتژی استفاده می‌کنند، جزء پیچیده‌ترین انواع برنامه‌ها هستند. این نوع برنامه‌ها خودشان را به صورتی تغییر می‌دهند که نمونه‌های جدید هیچ شباهتی به نمونه اصلی ندارد. این برنامه‌ها موتور کدگذاری ندارد و در هر انتقالی به‌طور خودکار کد منبع بدافزار تغییر می‌کند [۶].

برای تشخیص ناهنجاری الگوریتم‌های مختلفی بر مبنای یادگیری ماشین و داده‌کاوی استفاده می‌گردد که یک دسته از این روش‌ها از یادگیری جمعی استفاده می‌کنند. در یادگیری جمعی به‌جای یک دسته‌بند یا مدل از چندین دسته‌بند استفاده می‌گردد که در نهایت سبب ایجاد مدلی قوی‌تر می‌گردد. در زیر برخی از دلایل استفاده از یادگیری جمعی را برشمرده شده‌اند.

اولین دلیل، حجم زیاد داده‌هاست. در بسیاری از مسائل، تعداد نمونه به‌منظور تجزیه‌وتحلیل می‌تواند بیش‌ازحد بزرگ باشد تا بتواند به‌صورت مؤثر به‌وسیله یک دسته‌بند اداره شود، بنابراین آموزش یک دسته‌بند با چنین مقدار داده زیادی معمولاً عملی نیست [۷-۸]. دومین دلیل مربوط به شرایطی است که داده‌های خیلی کمی در اختیار داریم. در دسترس بودن مجموعه داده‌های آموزشی مناسب و کافی برای الگوریتم‌های طبقه‌بندی خیلی مهم است تا بتواند به‌صورت موفقیت‌آمیز از توزیع داده‌ها بیاموزد. در صورت عدم وجود داده‌های کافی، استفاده از روش‌های مبتنی بر یادگیری عمیق نیز توصیه می‌شود [۷]. سومین دلیل استفاده از یادگیری جمعی این است که این روش‌ها بر اساس رویکرد تقسیم و غلبه هستند. مسائلی وجود دارد که یک دسته‌بند، خیلی مشکل می‌تواند آن‌ها را حل کند. مخصوصاً اگر مرز تصمیم که داده‌های کلاس‌های مختلف را جدا می‌کند، خیلی پیچیده باشد یا در خارج از فضای عملکردی یک طبقه بند باشد [۹]؛ بنابراین نیاز است این مرز پیچیده تقسیم‌بندی شده و هر بخش توسط یکی از دسته‌بندهای شرکت‌کننده در یادگیری جمعی، جداسازی شود.

بدافزارها در محیط‌های مختلفی از جمله رایانه‌های شخصی، دستگاه‌های هوشمند و همراه، شبکه‌های رایانش ابری آسیب‌های مختلفی را ایجاد می‌کنند و تاکنون تحقیقات زیادی برای شناسایی و مقابله با آن‌ها در این محیط‌ها انجام شده است [۱۰-

۱۲]. روش‌های مبتنی بر هوش مصنوعی و یادگیری ماشین کاربرد فراوانی در این زمینه دارند [۱۳]. اخیراً استفاده از یادگیری جمعی برای شناسایی حملات و بدافزارها در محیط‌های مختلف بسیار موردتوجه قرار گرفته است. البته یادگیری جمعی به‌منظور حل مسئله دسته‌بندی، کاربردهای گسترده و بسیار متنوعی دارد و کاربرد این نوع یادگیری محدود به شناسایی حملات و بدافزارها نیست. در این الگوریتم‌ها معمولاً چندین یادگیرنده ضعیف با استفاده از داده‌های آموزشی ایجاد می‌گردند و این یادگیرندگان برای ایجاد مدلی قوی با هم ترکیب می‌شوند. الگوریتم AdaBoost یکی از الگوریتم‌های مهم یادگیری جمعی است که در این زمینه کاربرد فراوانی پیدا کرده است [۱۴-۱۵]. در واقع آدابوست یک متا الگوریتم است که به‌منظور ارتقاء عملکرد دیگر الگوریتم‌های یادگیری استفاده می‌شود. در این الگوریتم، طبقه بند هر مرحله جدید به نفع نمونه‌های غلط طبقه‌بندی‌شده در مراحل قبلی، تنظیم می‌گردد. آدابوست نسبت به داده‌های نویزی و پرت حساس است؛ ولی نسبت به مشکل بیش برازش، از بیشتر الگوریتم‌های یادگیری برتری دارد. الگوریتم یادگیری جمعی RandomForest یکی دیگر از این‌گونه الگوریتم‌هاست که کارایی خوبی در تشخیص بدافزارها دارد [۱۶-۱۸]. این الگوریتم یک از مهم‌ترین الگوریتم‌هایی است که از یک ساختار درختی جهت طبقه‌بندی و شناسایی بروی مجموعه‌ی داده‌ها در زمینه‌های خاص استفاده می‌گردد. این الگوریتم دقت بسیار بالایی به ازای مقادیر بالای داده‌ها دارد. در کل هدف اصلی این الگوریتم استفاده از ساختار درختی به‌منظور پیش‌بینی داده‌ها است. الگوریتم GradiantBoostedTrees سومین الگوریتمی است که در تشخیص و شناسایی بدافزارها در محیط‌های مختلف کارایی خوبی از خود نشان داده است [۱۹-۲۰]. گرادیان تقویتی یا گرادیان بوستینگ، یک روش یادگیری ماشین برای مسائل رگرسیون و طبقه‌بندی است. مدل گرادیان تقویتی ترکیبی خطی از یک سری مدل‌های ضعیف است که به‌صورت تناوبی برای ایجاد یک مدل نهائیِ قوی ساخته شده است؛ مانند دیگر روش‌های تقویتی (بوستینگ)، گرادیان تقویتی (گرادیان بوستینگ) ترکیبی خطی از یک سری از مدل‌های ضعیف برای ایجاد یک مدل قوی و کارآمد است.

اخیراً یادگیری عمیق به‌منظور شناسایی بدافزارها موردتوجه قرار گرفته و کارایی خوبی از خود نشان داده است [۲۱-۲۲]. با توجه به اهمیت و کاربرد روزافزون رایانش ابری، در این تحقیق، شناسایی بدافزارها و حملات مربوطه در محیط رایانش ابری با استفاده از یادگیری جمعی موردتوجه قرار گرفته است. ساختار و الگوریتم پیشنهادی ما بر مبنای ترکیب سه الگوریتم یادگیری جمعی نام برده شده است و در واقع یک الگوریتم یادگیری جمعی است که خود از سه الگوریتم یادگیری جمعی مهم و کارا در این زمینه بهره می‌برد.

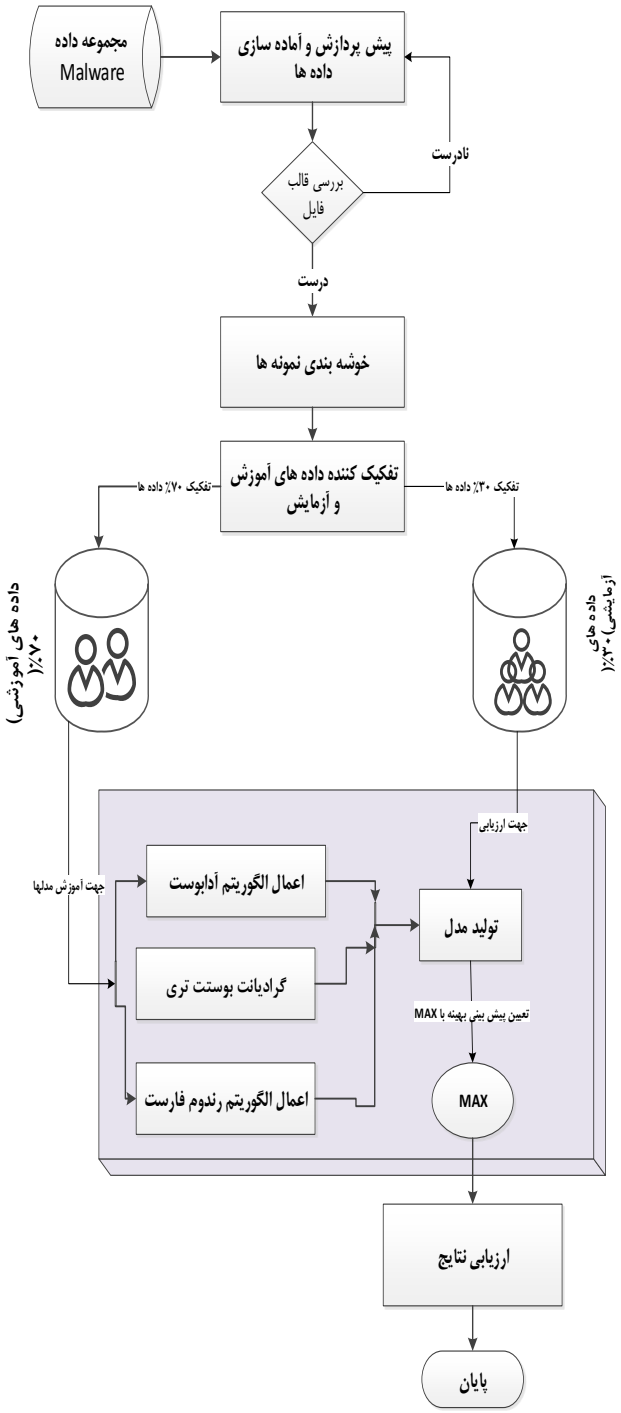
۳- روش پیشنهادی

ایده کلی روش پیشنهادی بهره‌گیری از چند سیستم یادگیرنده جمعی است. به‌نحوی‌که در لایه اول دسته‌بندی سه دسته‌بند جمعی قوی به کار می‌رود سپس از مکانیزم رأی‌گیری برای تصمیم‌گیری نهایی استفاده می‌گردد. همچنین قبل از انجام دسته‌بندی، داده‌های ورودی، خوشه‌بندی می‌شوند تا به‌منظور آموزش این سه دسته‌بند داده‌های بهتری داشته باشیم زیرا هرکدام از دسته‌بندهای جمعی از یادگیرنده‌های ضعیفی استفاده می‌کنند و لازم است داده‌های باکیفیت‌تری را برای آموزش خود دریافت کنند. از طرف دیگر خوشه‌بندی مشکل کلاس‌های نامتوازن را حل می‌کند زیرا درصورتی‌که تعداد نمونه‌های یک یا چند کلاس از کلاس‌های دیگر بیشتر باشد، دسته‌بند تمایل بیشتری به کلاس‌های اکثریت پیدا می‌کند، زیرا قوانینی که این نمونه‌ها را به‌درستی پیش‌بینی می‌کنند به‌درستی وزن‌دهی شده‌اند درحالی‌که قوانین خاصی که نمونه‌های کلاس اقلیت را پیش‌بینی می‌کنند عموماً نادیده گرفته می‌شوند و در واقع به‌صورت نویز با آن‌ها برخورد می‌شود و در نتیجه نمونه‌های کلاس اقلیت به‌اشتباه کلاسه‌بندی خواهند شد؛ بنابراین انجام خوشه‌بندی قبل از دسته‌بندی سبب افزایش توان دسته‌بند می‌گردد. در این روش خوشه‌های مختلف یک کلاس همگی در نهایت عضو همان کلاس محسوب می‌شوند.

برای اجرای روش پیشنهادی، در ابتدا مجموعه داده‌های بدافزارها برای استفاده در نرم‌افزار رپیدماینر یا متلب پیش‌پردازش می‌شوند. به دلیل زیاد بودن حجم اطلاعات ممکن است لازم به حذف نمونه‌های تکراری در مجموعه داده باشیم. پس‌ازآن با استفاده از روش kامین ارزیابی، داده‌ها به k قسمت برای ایجاد داده‌های آموزش و آزمایش، تقسیم می‌شود. در این روش داده‌ها به k بلوک تقسیم می‌شوند. در هر تکرار یکی از k بلوک به‌عنوان داده آزمون و مابقی به‌عنوان داده آموزش در نظر گرفته می‌شوند. مطابق شکل (۲) به‌طورکلی فرآیند تشخیص حملات بدافزارها در ایانش ابری عبارت‌اند از: در مرحله اول گره‌ها داده‌هایی را تولید می‌کنند یا مجموعه‌ای از داده‌ها توسط مهاجمین مختلف به سیستم وارد می‌شود. در مرحله اول این داده‌ها به سیستم تزریق می‌شوند. در مرحله دوم یک پیش‌پردازش کلی بروی داده‌ها و نمونه‌های واردشده صورت می‌گیرد. با پیش‌پردازش داده‌ها، مقادیر پرت از بین خواهند رفت. در این مرحله با کمک خوشه‌بندی، داده‌ها را تفکیک نموده و مشخص می‌شود که داده‌های تزریق‌شده در چه دسته‌هایی

تقسیم‌بندی شده‌اند؛ بنابراین خروجی این مرحله که تعیین گروه برای داده‌ها است به مرحله بعد که تفکیک داده‌ها می‌باشد، ارسال می‌گردد.

در مرحله بعد داده‌های آموزشی جهت تولید مدل و داده‌های آزمایشی برای ارزیابی روش پیشنهادی و تشخیص ناهنجاری‌ها از نمونه‌های جدید واردشده تفکیک می‌شوند. در نهایت بر روی داده‌های مسئله، الگوریتم ترکیبی پیشنهادی اعمال‌شده و ناهنجاری‌ها تشخیص داده می‌شوند. استراتژی عملکرد روش یادگیری جمعی پیشنهادی در این تحقیق مبتنی بر رأی‌گیری است. مجموعه داده‌های آموزشی به هرکدام از الگوریتم‌های آدابوست[1]، رندوم فارست[2] و درخت گرادینت بوستد[3] وارد شده و مدل‌های این الگوریتم‌ها تولید می‌شوند. سپس داده‌های تست به مدل‌های الگوریتم‌های ذکر شده، وارد می‌گردد و هر الگوریتم برای هر نمونه تست یک پاسخی را صادر می‌کند. هر پاسخ بیانگر این است که یک نمونه آیا حمله تشخیص داده شده است یا خیر. پاسخ‌های کلیه الگوریتم‌ها به هسته رأی‌گیری[4] وارد می‌شوند. هسته رأی‌گیری بر اساس استراتژی ماکسیمم[5] تصمیم می‌گیرد که چه پاسخی در نهایت به خروجی ارسال شود. سه الگوریتم استفاده‌شده در مدل پیشنهادی، خود دسته‌بندهای قوی هستند که به‌تنهایی در مسائل مختلف از جمله مسئله تشخیص بدافزار و حملات کارایی خوبی دارند. حال وقتی‌که مکانیزم رأی‌گیری روی خروجی‌های این سه دسته‌بند اعمال می‌شود، نتیجه بهتری را تولید خواهد کرد و میزان خطای بسیار کمتری خواهد داشت. برای انجام رأی‌گیری از تعداد فردی از دسته‌بندها استفاده می‌شود که در اینجا از سه دسته‌بند استفاده شده است تا همواره برای هر نمونه داده، اجماع وجود داشته باشد و بتوان با استفاده از نظر اکثریت، برچسب هر داده ورودی را مشخص کرد. ازآنجایی‌که سه دسته‌بند به‌صورت متوالی اجرا می‌شوند، بر اساس قوانین تحلیل الگوریتم‌ها، پیچیدگی محاسباتی دسته‌بند ایجاد شده توسط مکانیزم رأی‌گیری از پیچیدگی محاسباتی پیچیده‌ترین دسته‌بند استفاده شده در این مجموعه، بیشتر نخواهد شد؛ یعنی حلقه یا تکرار جدیدی اضافه نمی‌شود که سبب بالا رفتن پیچیدگی محاسباتی شود، اگرچه انتظار می‌رود زمان آموزش نسبت به هریک از سه روش پایه افزایش یابد، چون به‌جای یک دسته‌بند، سه دسته‌بند باید آموزش ببینند. در واقع مدل دسته‌بندی نهایی یک یادگیری جمعی است که خود روی نتیجه یادگیرنده‌های جمعی دیگر انجام می‌شود. همان‌طور که در شکل (۲) مشخص است داده‌های آموزشی از یک طرف برای ساخت مدل ترکیبی پیشنهادی وارد سیستم می‌شوند و از طرف دیگر داده‌های آزمایشی بعد از ساخت مدل برای ارزیابی مدل نهایی به کار می‌روند. برای ارزیابی مدل، از معیارهای مختلفی استفاده شده است.



شکل (۲). فلوچارت روش پیشنهادی

رابطه‌های زیر به ترتیب بیانگر معیارهای دقت، میزان فراخوانی، صحت و میزان خطا هستند که در اینجا برای ارزیابی و مقایسه مدل پیشنهادی از آن‌ها استفاده شده است.

[1] Adaboost
[2] Random Forest
[3] Gradient Boosted Trees
[4] voting
[5] maximum

رابطه (1) $$precision = \frac{TP}{TP+FP}$$

رابطه (2) $$Recall = \frac{TP}{TP+FN}$$

رابطه (3) $$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

رابطه (4) $$Error = 100 - \frac{TP+TN}{TP+TN+FP+FN}$$

۴- ارزیابی روش پیشنهادی

روش پیشنهادی در این مقاله با استفاده از شبیه‌ساز رپیدماینر نسخه ۷٫۳ پیاده‌سازی شده است. همچنین در جدول (۱) مشخصات مربوط به سیستمی که پیاده‌سازی روش پیشنهادی و روش‌های قبلی و همچنین ارزیابی نتایج در آن انجام شده، نشان داده شده است.

جدول (۱). مشخصات سیستم جهت شبیه‌سازی و ارزیابی نتایج

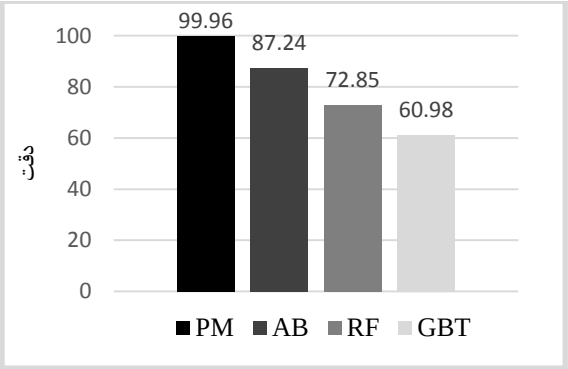
سخت‌افزار / نرم‌افزار	مشخصات
سیستم‌عامل	ویندوز 10
نوع سیستم‌عامل	سیستم‌عامل 64 بیتی
حافظه RAM	12 گیگابایت
پردازنده	پردازنده اینتل (Core™i7 CPU)
ابزار مدل‌سازی	نرم‌افزار داده‌کاوی رپیدماینر
منبع داده استفاده شده	UNSW-NB15 data set

منبع داده مورداستفاده شامل دو قسمت داده‌های آموزشی و داده‌های آزمون می‌باشد. این دیتاست جهت تشخیص حملات Malware در رایانش ابری استفاده شده است. از داده‌های آموزشی برای تولید مدل و از داده‌های آزمایشی برای تست و ارزیابی مدل پیشنهادی استفاده می‌گردد. بسته‌های شبکه خام مجموعه داده UNSW-NB15 توسط ابزار XIA PerfectStorm در آزمایشگاه سایبر محدوده مرکز امنیت سایبر استرالیا (ACCS) برای تولید هیبریدی از فعالیت‌های عادی واقعی مدرن و رفتارهای حملات مصنوعی معاصر ایجاد شده است. این مجموعه داده دارای نه نوع حمله است، یعنی Fuzzers، Analysis، Backdoors، DoS، Exploits، Generic، Reconnaissance، Shellcode و Worms. از ابزارهای Argus، Bro-IDS استفاده شده و ۴۹ ویژگی با برچسب کلاس ایجاد شده است. نتایج شبیه‌سازی الگوریتم‌ها به‌صورت تک‌به‌تک بر روی منبع داده‌ها ارزیابی شده و مقایسه می‌شود؛ بنابراین، در این قسمت نتایج به‌دست‌آمده در روش پیشنهادی را با سایر روش‌های جمعی مورد مقایسه قرار می‌دهیم. در جدول (۲) بر اساس معیارهای مختلف، روش پیشنهادی در مقایسه با روش‌های دیگر نشان داده شده است. در این جدول برای هر روش یک نام اختصاری مشخص شده است، که در ادامه برای سهولت در ارائه نتایج، از این نام‌ها استفاده خواهیم کرد.

جدول (۲). مقایسه نتایج روش پیشنهادی با روش‌های ارائه شده قبلی

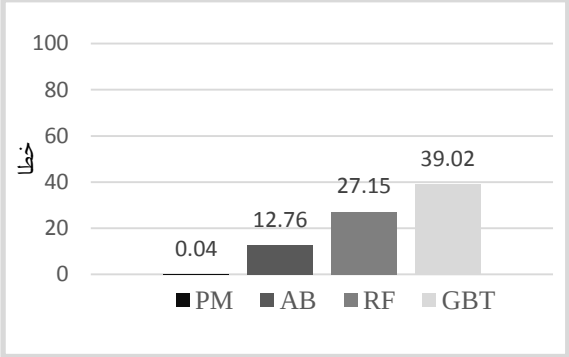
معیار	روش پیشنهادی (PM)	Ada Boost (AB)	Random Forest(RF)	Gradient Boosted Trees(GBT)
دقت	۹۹/۹۶	۸۷/۲۴	۷۲/۸۵	۶۰/۹۸
خطا	۰/۰۴	۱۲/۷۶	۲۷/۱۵	۳۹/۰۲
میزان صحت	۹۹/۹۷	۸۷/۶۸	۷۹/۴۸	۳۰/۴۹
میزان فراخوانی	۹۹/۹۵	۸۹/۵۴	۷۷/۷۴	۵۰/۰۰

همان‌طور که مشاهده می‌شود روش پیشنهادی از لحاظ میزان دقت، خطا، میزان صحت و میزان فراخوانی، نسبت به سایر روش‌ها بهینه‌تر می‌باشد. در شکل (۳) مقایسه دقت روش پیشنهادی نسبت به سایر روش‌هایی که تاکنون مطرح شده است نشان داده شده است. در این نمودار، محور افقی دسته‌بندها را معرفی می‌کند و محور عمودی نشان‌دهنده دقت است.



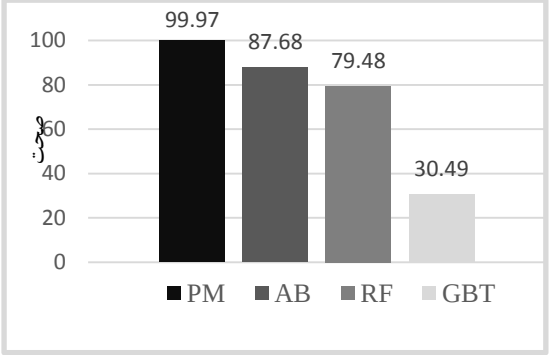
شکل (۳). مقایسه دقت روش پیشنهادی با سایر روش‌ها

همان‌طور که مشاهده می‌شود دقت روش پیشنهادی به‌منظور تشخیص حملات در شبکه‌های رایانش ابری برابر با ۹۹٫۹۶٪ بوده است که این میزان دقت بسیار قابل‌قبول بوده و نسبت به سایر روش‌هایی که تاکنون انجام شده‌اند، بهتر است. استفاده از سه دسته‌بند قوی در مدل پیشنهادی و همچنین متوازن‌سازی کلاس‌ها توسط خوشه‌بندی، دلایل اصلی دقت بالای روش پیشنهادی هستند. در شکل (۴) مقایسه خطا روش پیشنهادی نسبت به سایر روش‌هایی که تاکنون مطرح شده است نشان داده شده است. در این شکل نیز محور افقی معرفی کننده دسته‌بندهاست و محور عمودی خطا در دسته‌بندی را مشخص می‌کند.

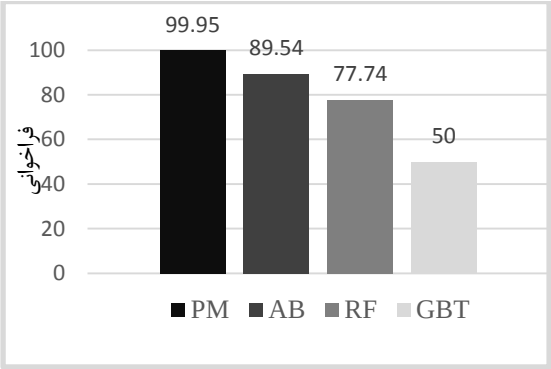


شکل (۴). مقایسه خطای روش پیشنهادی نسبت به سایر روش‌ها

همان‌طور که مشاهده می‌شود در روش پیشنهادی میزان خطای تشخیص حملات در شبکه‌های رایانش ابری برابر ۴٪ می‌باشد که این میزان خطا بسیار ناچیز بوده و در جهت توسعه این سیستم‌ها نمی‌تواند اختلالی وارد نماید. اصولاً وقتی دقت یک روش دسته‌بندی بالا باشد، با توجه به روابط ارائه شده در انتهای بخش سوم، انتظار می‌رود که خطا کم باشد؛ و این آزمایش برای تأکید بیشتر و مقایسه خطاها انجام گرفت که در نتیجه برتری روش پیشنهادی از این نظر هم تأیید می‌شود.



شکل (۵). مقایسه صحت روش پیشنهادی نسبت به سایر روش‌ها



شکل (۶). مقایسه فراخوانی روش پیشنهادی نسبت به سایر روش‌ها

نمودارهای ستونی شکل (۵) و شکل (۶) به ترتیب میزان صحت و فراخوانی هرکدام از روش‌ها در مسئله دسته‌بندی بدافزارها در کلاس‌های مختلف را نشان می‌دهند. در این نمودارها، محورهای افقی، روش‌های دسته‌بندی را مشخص کرده‌اند و محورهای عمودی به ترتیب نشان‌دهنده صحت و فراخوانی هستند. با توجه به شکل (۵)، صحت روش پیشنهادی برابر با ۹۹/۹۷ و با توجه به شکل (۶) فراخوانی روش پیشنهادی برابر با ۹۹/۹۵ است که از نظر صحت و فراخوانی، روش پیشنهادی برتری قابل‌توجهی نسبت به سایر روش‌ها، دارد. علت کارایی بهتر روش پیشنهادی با سایر روش‌های مبتنی بر یادگیری جمعی، استفاده از خوشه‌بندی به‌منظور متوازن‌سازی کلاس‌هایی است که در حالت عادی تأثیر منفی بر روی دسته‌بندی ایجاد می‌کنند. همچنین استفاده از یادگیری جمعی بر روی یادگیرنده‌های جمعی قوی که هریک خود از یادگیرنده‌های ضعیف زیادی بهره می‌برند سبب شده است که یک مدل ترکیبی قوی برای تشخیص بدافزارها به دست آید.

۵- نتیجه‌گیری و کارهای آینده

امروزه شبکه‌های رایانش ابری در بسیاری از کاربردهای مختلف مورداستفاده قرار می‌گیرند. در این شبکه‌ها، مشکلات، چالش‌ها و مخاطراتی وجود دارد که موجب ایجاد اختلال در روند شبکه و کاربردهای آن شده است. مهم‌ترین چالشی که امروزه محققین با آن مواجه هستند، حمله مهاجمین با استفاده از حملات مختلف به ارتباطات شبکه‌ای، مرکز داده، خدمات مشتری و غیره است. در این مقاله، با استفاده از خوشه‌بندی به‌منظور تفکیک داده‌ها و همچنین یادگیری جمعی، ناهنجاری‌ها در شبکه رایانش ابری تشخیص داده می‌شوند. همچنین می‌توان برای هریک از شاخص‌های داده، امتیازی در نظر گرفت تا معیارهای مناسب برای شناسایی ناهنجاری‌ها انتخاب گردند. در این صورت علاوه برافزایش دقت تشخیص، سرعت محاسبات افزایش‌یافته و مصرف انرژی در شبکه کاهش می‌یابد. با توجه به ارزیابی‌های انجام شده، مشاهده گردید که روش پیشنهادی نسبت به سایر روش‌ها از نظر دقت، میزان خطا، صحت و فراخوانی بهتر عمل می‌کند و می‌توان به این روش اعتماد نمود. بر اساس آزمایش‌های انجام شده، روش پیشنهادی بهبود دقت ۱۲,۷۲٪ را نشان می‌دهد، درحالی‌که افزایش چندانی در پیچیدگی محاسباتی ایجاد نمی‌کند. به‌منظور بهبود بیشتر کارایی در شناسایی بدافزارها در محیط رایانش ابری، می‌توان از دسته‌بندهای مبتنی بر یادگیری عمیق نیز در ساختار پیشنهادی بهره گرفت. محاسبه ریسک استفاده از یک نرم‌افزار به کمک تحلیل ایستا و پویای آن در محیط رایانش ابری، به‌منظور اعلام هشدار در استفاده از نرم‌افزارهای مخرب، یکی دیگر از تحقیقاتی است که می‌تواند در آینده موردتوجه قرار گیرد.

ی همگرایی و

۶- مراجع

[1] S. Naval, V. Laxmi, M. Rajarajan, M. S. Gaur, & M. Conti, “Employing Program Semantics for Malware Detection,” *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 12, pp. 2591-2604, Dec. 2015.

[2] Z. Bazrafshan, H. Hashemi, S. M. H. Fard, & A. Hamzeh, “A survey on heuristic malware detection techniques,” in IKT 2013 5th Conference on Information, 2013.

[3] A. Damodaran, F. D. Troia, C. A. Visaggio, T. H. Austin, & M. Stamp, “A comparison of static, dynamic, and hybrid analysis for malware detection,” *J. comput. virol. hacking tech.*, vol. 13, no. 1, pp. 1-12, 2017.

[4] M. Ahmadi, A. Sami, H. Rahimi, & B. Yadegari, “Malware detection by behavioural sequential patterns,” *Comput. fraud secur.*, vol. 2013, no. 8, pp. 11-19, 2013.

[5] H. Darabian, A. Dehghantanha, S. Hashemi, S. Homayoun, & K.-K. R. Choo, “An opcode based technique for polymorphic Internet of Things malware detection: An OpCode-Based Technique for Polymorphic Internet of Things Malware Detection,” *Concurr. Comput.*, vol. 32, no. 6, p. e5173, 2020.

[6] B. B. Rad, M. Masrom, & S. Ibrahim, “Opcodes histogram for classifying meta-morphic portable executables malware,” in *2012 International Conference on e-Learning and e-Technologies in Education (ICEEE)*, IEEE, pp. 209-213, 2012.

[7] W. Hardy, L. Chen, S. Hou, Y. Ye, & X. Li, “DL 4 MD: A deep learning framework for intelligent malwarc detection,” in *Inel Conf. Data Mining*, CSREA Press, pp. 61-67, 2016.

[8] L. Yu, S. Wang, & K. K. Lai, “Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm,” *Energy Econ.*, vol. 30, no. 5, pp. 2623-2635, 2008.

[9] S. Almarri & P. Sant, “Optimised Malware Detection in Digital Forensics,” *Int. j. netw. secur. appl.*, vol. 6, no. 1, pp. 1-15, 2014.

[10] M. Deypir, “Entropy-based security risk measurement for Android mobile applications,” *Soft Comput.*, vol. 23, no. 16, pp. 7303-7319, 2019.

[11] Ram Mahesh Yadav, “Effective analysis of malware detection in cloud computing,” Computers & Security, vol. 83, pp. 14-21, 2019.

[12] M. Ghasabi, M. Deypir, & E. Mahdipour, "A New Algorithm Based on Hellinger Distance for Mitigation of DDoS Attacks in Software Defined Networks," Journal of Electronical & Cyber Defence Vol. 5, No. 4, 2017. (In Persian)

[13] N. Usman, S. Usman, F. Khan, M. A., Jan, A. Sajid, M. Alazab, & P. Watters, "Intelligent dynamic malware detection using machine learning in IP reputation for forensics data analytics," Future Generation Computer Systems, 118, vol. 118, pp. 124-141, 2021.

[14] A. Shahraki, M. Abbasi, & Ø. Haugen, “Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost,” *Eng. Appl. Artif. Intell.*, vol. 94, no. 103770, p. 103770, 2020.

[15] L. Pallippattu Mathai, "Malware Detection on Android using Adaboost Algorithm," Doctoral dissertation, Dublin, National College of Ireland, 2021.

[16] F. C. Garcia & F. P. Muga II, "Random forest for malware classification," arXiv preprint arXiv:1609.07770, 2016.

[17] C. Galen & Steele, R. “Performance Maintenance Over Time of Random Forest-based Malware Detection Models,” 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pp. 0536-0541, IEEE, October 2020.

[18] S. Joshi, H. Upadhyay, L. Lagos, N. S. Akkipeddi, & V. Guerra, "Machine learning approach for malware detection using random forest classifier on process list data structure," In Proceedings of the 2nd International Conference on Information System and Data Mining, pp. 98-102, April 2018.

[19] H. D. Pham, T. D. Le, & T. N. Vu, "Static PE malware detection using gradient boosting decision trees algorithm," In International Conference on Future Data and Security Engineering, pp. 228-236, Springer, Cham, November 2018.

[20] C. Galen & R. Steele, "Empirical Measurement of Performance Maintenance of Gradient Boosted Decision Tree Models for Malware Detection," In 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 193-198, IEEE, April 2021.

[21] J. E. L. Abdelkhalki, M. B. Ahmed, & A. A. Boudhir, “Image malware detection using deep learning,” *IJCNIS*, vol 12, no 2, 2020.

[22] D. Tian, Q. Ying, X. Jia, R. Ma, C. Hu, & W. Liu, “MDCHD: A novel malware detection method in cloud using hardware trace and deep learning,” Computer Networks, vol 198, bl 108394, 2021.