شناسایی بدافزار فراریخت مبتنی بر نرخ تکرار کدهای عملیاتی و ثبات‌ها با استفاده از روش‌های همبستگی و فاصله

هادی گلباغی[1]، محمد فتحی[2]، فرشته کیاست[3]

[1] کارشناسی ارشد، مرکز آپا دانشگاه کردستان، سنندج
h.golbaghi@uok.ac.ir

[2] دانشیار، مرکز آپا دانشگاه کردستان، سنندج
mfathi@uok.ac.ir

[3] کارشناسی ارشد، مرکز آپا دانشگاه کردستان، سنندج
f.kiasat@uok.ac.ir

چکیده

در طراحی بدافزارها تلاش بر این است تا از روش‌هایی برای جلوگیری از شناسایی شدن استفاده شود. یکی از انواع بدافزارها، بدافزار فراریخت است که در هر انتشار ساختار کد خود را تغییر می‌دهد. روش‌های شناسایی بدافزار در مواجه با بدافزار به دلیل جایگزین‌کردن دستورات مشابه کد اصلی، رمزنگاری ساختار کد بدافزار یا درج کد زائد دچار شکست می شوند و در برخی موارد، سربار محاسباتی بالا، دقت شناسایی و کارایی ضعیف روش‌ها، آن‌ها را دچار چالش می‌کند. روش پیشنهادی این مقاله، شناسایی از طریق تحلیل ایستای نرخ تکرار کدهای عملیاتی و ثبات‌ها مبتنی بر دو معیار ریاضی بیشینه ضریب همبستگی و کمینه معیار فاصله است. به منظور ارزیابی این روش‌ها، آزمایش‌هایی در حالات مختلف بر روی ۵۰، ۱۰۰، ۱۵۰، ۲۰۰، ۲۵۰ و ۴۴۰ فایل متشکل از فایل‌های سالم و چهار خانواده بدافزار فراریخت از ویروس‌ها و کرم‌های *G2, MPCGEN, MWOR, NGVCK* انجام می‌گیرد. نتایج آزمایش‌ها نشان می‌دهد که روش‌های پیشنهادی نسبت به روش‌های قبلی، به دقت نسبتاً بالایی در شناسایی بدافزارهای فراریخت می‌رسند.

کلمات کلیدی : بدافزار، فراریختی، کدهای عملیاتی، کمینه فاصله، ضریب همبستگی

۱- مقدمه

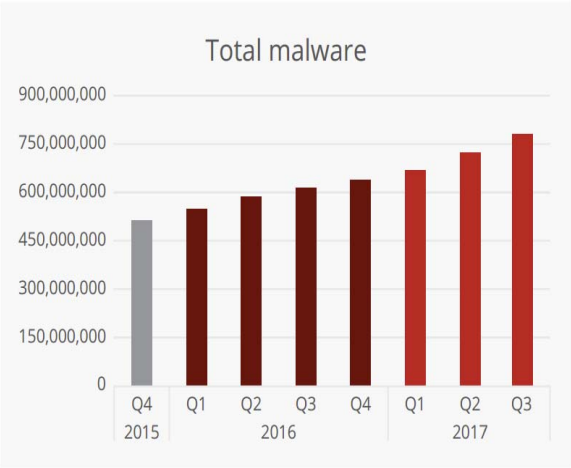
با گسترش روز افزون فناوری‌های نوین و راه پیدا کردن اینترنت در زندگی انسان‌ها، زمینه برای نفوذ و آلوده شدن سیستم‌های کامپیوتری فراهم شده است. تعاریف مختلفی از بدافزار[۱] و بدافزارهای مخرب[۲] وجود دارد. بدافزارها برنامه‌های کامپیوتری هستند که هدف آن‌ها آسیب‌رسانی به سیستم‌های کامپیوتری است [۱]. هر کدی که به صورت عمدی باعث آسیب رسانی یا متوقف کردن عملیات روتین سیستم شود را بدافزار می‌گویند [۲].

بدافزارها تعمداً برای انجام عملیات بدون مجوز، مخرب یا نامطلوب طراحی می‌شوند [۳]. بدافزارها علاوه بر اینکه با عملیات خرابکارانه به منابع و اطلاعات سیستم‌ها دسترسی پیدا می‌کنند، به سرویس‌دهی و فعالیت‌های سیستم‌ها آسیب رسانده و با سرقت غیرمجاز اطلاعات شخصی افراد هزینه‌های روحی و روانی زیادی در کنار ضررهای مالی در زمینه افشای حریم خصوصی وارد می‌کنند [۴]. به همین دلیل، مسئله شناسایی بدافزارها در دو دهه اخیر مورد توجه شرکت‌های ضدبدافزار و محققان بوده است. بدافزارها را بر اساس هدف نویسندگان و نوع حملات آن‌ها به دسته‌های مختلفی از قبیل ویروس[۳]، کرم[۴]، تروجان[۵]، دربپشتی[۶]، جاسوس افزار[۷]، باج افزار[۸]، رد گم کن[۹] و غیره تقسیم کرده‌اند [۵].

به طور متوسط، در هر حادثه جرایم سایبری ۱۹۷ دلار از دست می‌رود [۶]. در شکل (۱)، آمار مجموع تعداد بدافزارها و رشد آن‌ها از سه ماهه چهارم سال ۲۰۱۵ تا سه ماهه سوم سال ۲۰۱۷ نشان داده شده است. طبق این آمار، تولید بدافزارهای جدید در سه ماهه سوم سال ۲۰۱۷ بالاتر از ۵۷ میلیون مورد بوده که بالاترین میزان ثبت شده تا کنون است و نسبت به دوره قبل دارای رشد ۱۰ درصدی نیز می باشد [۷]. با توجه به این گزارش‌ها می‌توان پی برد که اصولاً شناسایی بدافزارها کاری پیچیده و سخت است. علت اینست که نویسندگان بدافزارها به طور مداوم روش‌های جدیدی برای مقابله با روش‌های کشف موجود طراحی می‌کنند.

یکی از گونه‌های بدافزار، بدافزارهای فراریخت[۱۰] است. ساختار کد بدافزارهای فراریخت، در هر تکثیر با استفاده از فنون مختلف مبهم‌سازی کد[۱۱] تغییر می‌یابند اما عملکرد اصلی آن‌ها حفظ می‌شود و این مسئله، شناسایی آن‌ها را دشوار و پیچیده می‌کند [۳]. روش‌های مختلفی برای شناسایی بدافزارهای فراریخت پیشنهاد وجود دارند. به عنوان یک دسته‌بندی کلی می‌توان روش‌های کشف بدافزار را به دو دسته تحلیل ایستا[۱۲] و پویا[۱۳] تقسیم کرد. روش‌های ایستا محبوب‌تر و رایج‌ترند؛ زیرا بدون نیاز به اجرای بدافزار و صرفاً با تحلیل ساختار کد بدافزار، اقدام به کشف آن‌ها می‌نمایند [۴]. یکی از روش‌های شناسایی بدافزار مبتنی بر تحلیل ایستا که بسیار مورد استفاده است، روش‌های مبتنی بر امضا[۱۴] هستند [۸]. در مقابل، روش‌های پویا با اجرای بدافزار و جمع‌آوری اطلاعاتی از اجرای بدافزار سعی می‌کنند بدافزار را شناسایی نمایند [۹].

روش پیشنهادی در این مقاله مبتنی بر تحلیل ایستای تفاوت نرخ تکرار کدهای عملیاتی[۱۵] و ثبات‌ها[۱۶] در خانواده‌های مختلف بدافزار فراریخت و فایل‌های سالم، اطلاعاتی را جمع‌آوری می‌نماید. کشف گونه جدید بدافزار فراریخت و جدا کردن آن از فایل‌های سالم، با استفاده از معیارهای بیشینه ضریب همبستگی[۱۷] و کمینه فاصله[۱۸] صورت خواهد گرفت. روش پیشنهادی با تأکید بر روی پالایش و انتخاب صحیح کدهای عملیاتی در صدد افزایش کارایی است و در شمارش کدهای عملیاتی، با همسان در نظر گرفتن دستورات مشابه، سعی دارد در مقابل روش مبهم‌سازی جانشینی دستورات مشابه مقاوم باشد.



شکل (۱) : مجموع بدافزارها در پایگاه داده آزمایشگاه مک‌آفی [۷]

برای نشان دادن عملکرد روش پیشنهادی در موقعیت‌های عملی و ارزیابی آن، آزمایش‌هایی در حالات مختلف بر روی ۵۰، ۱۰۰، ۱۵۰، ۲۰۰، ۲۵۰ و ۴۴۰ فایل متشکل از فایل‌های سالم و چهار خانواده بدافزارهای فراریخت از ویروس‌ها و کرم‌های *G2, MPCGEN, MWOR, NGVCK* انجام می‌شود. نتایج آزمایش‌ها نمایانگر دقت بالای روش پیشنهادی است.

در ادامه این مقاله دربخش دوم کارهای مرتبط مورد بررسی قرار می‌گیرد. در بخش سوم روش پیشنهادی تشریح می‌شود. بخش چهارم به ارزیابی و تفسیر نتایج اختصاص دارد. در نهایت بخش پنجم نیز به نتیجه‌گیری وکارهای آینده می‌پردازد.

۲- کارهای مرتبط

در سالیان اخیر روش‌های متعددی برای کشف بدافزارهای فراریخت پیشنهاد شده‌اند که دارای نقاط قوت و ضعفی هستند. برخی از روش‌ها [۴، ۱۰] به دلیل سربار محاسباتی بالا، زمانی در مقابل ویروس‌ها موثر هستند که زمان کافی برای شناسایی داشته و ضدبدافزار و بانک اطلاعاتی آن با نرخ سریعی به‌روزرسانی شوند. در غیر این صورت، تاثیر مثبت آن‌ها کاهش پیدا می‌کند [۱۱].

روش پیشنهادی ران‌وال و همکاران [۹] از جمله روش‌های اندازه‌گیری شباهت بر اساس گراف کدهای عملیاتی است و کارآمدتر از روش ارائه شده توسط آندرسون و همکاران [۱۲] عمل می‌کند. این روش با دریافت یک فایل اجرایی، رشته کدهای عملیاتی را استخراج و از روی آن گراف وزن‌دار را می‌سازد؛ اما به جای استفاده از هسته گراف‌ها، به طور مستقیم گراف‌های کد عملیاتی را مقایسه می‌کند. بر اساس نتایج، روش پیشنهادی دارای دقت نسبتا خوبی بوده و کارایی بهتری نسبت به روش‌های قبلی دارد. البته اگر از روش جایگزینیِ دستوراتِ مشابه استفاده شود، تشخیص فراریختی توسط روش پیشنهادی این مقاله دشوار خواهد بود.

گامال محمد و نورافیدا بنتی در [۱۳] قالبی را پیشنهاد می‌کنند که منجر به ایجاد رویکردی جدید بر اساس روش‌های مبتنی بر امضا و مبتنی بر رفتار رشته محور برای بهبود شناسایی بدافزارهای فراریخت شده است. شناسایی با استفاده از مجموعه داده‌های استاندارد از نمونه بدافزارهای شناخته شده به صورت فرمت

رشته‌ای، توابع و پارامترهای مختلف انجام می‌شود. نتایج نشان می دهد که بخش‌های پر خطر کدها و فایل‌های بدافزار از قطعه کدهایی هستند که به همراه دستورات سالم به کد اصلی تزریق شده‌اند. در نتیجه، این روش شناسایی بدافزارهای ناشناخته را نیز تسهیل می‌کند و سرعت و دقت را با کاهش پیچیدگی محاسباتی در زمان تشخیص بدافزار و کاهش حافظه مصرفی افزایش می‌دهد.

روش پیشنهادی کانفورما و همکاران [۱۴] به معرفی فنون شناسایی تکیه می‌کند که مبتنی بر فرض وجود یک اثر جانبی مشترک بین بسیاری از موتورهای فراریخت می‌باشد. این روش برای حدود ۱۰۰۰ برنامه، مورد آزمایش و بررسی قرار داده شده و بر اساس نتایج آن به طور دقیق ویروس‌های فراریخت و غیر فراریخت را دسته‌بندی می‌کند. از معایب این است که اگر از روش جایگزینیِ دستوراتِ مشابه یا اضافه کردن کد زائد به کد بدافزار و یا رمزنگاری بدنه کد بدافزار استفاده شود، توزیع دستورات تکراری تغییر می‌کند و روش پیشنهادی دچار شکست در شناسایی خواهد شد.

رویکرد مهرا و همکاران [۱۵] بر روی شناسایی و طبقه‌بندی بدافزارهای فراریخت بر اساس خانواده‌های آن‌ها تمرکز دارد. روش پیشنهادی به این صورت است که گراف جریان کنترلی رسم شده و گراف فراخوانی *API* ها ایجاد می‌شود. این رویکرد هر بدافزار فراریخت را براساس ویژگی‌های خانواده‌شان که از هیستوگرام و فرمول اندازه‌گیری کای دو، که بر اساس تحلیل پویا است، طبقه‌بندی می‌کنند. در این مقاله، دقت در الگوریتم‌های مختلف طبقه‌بندی از ۸۹ تا ۹۹/۱۰ درصد به‌دست آمده است.

به اعتقاد محمد بن خمس و همکاران [۱۶] هنوز هم دستیابی به دقت کامل و کارایی مناسب برای شناسایی بدافزارهای فراریخت یک چالش محسوب می‌شود. روش پیشنهادی این مقاله ویژگی‌های تغییر داده نشده در ساختار بدافزار را برای استفاده در فرایند شناسایی با استفاده از ماشین بردار پشتیبانی استخراج می‌کند. خصوصیات *n-gram* به‌طور مستقیم از ساختار باینری بدافزار استخراج شده، که این خصوصیات به عنوان امضا در نظر گرفته می‌شوند. این خصوصیات مقادیر قابل توجهی از تعداد خصوصیات انتخاب *n-gram* در حالت اصلی را کاهش می‌دهد. این روش ترکیبی از استخراج امضا *n-gram* و ماشین بردار پشتیبانی است. نتایج ارزیابی روش پیشنهادی برای شناسایی بدافزارهای فراریخت نشان می‌دهد که این روش قادر است دقتی در حدود ۹۹ درصد و نرخ منفی کاذب پایینی داشته باشد.

۳- روش پیشنهادی

با تحلیل و بررسی روش‌های موجود در حوزه شناسایی بدافزار فراریخت و دقت به نقاط قوت و ضعف آن‌ها نشان می‌دهد که موتورهای فراریختی همه موارد را در ساختار کد بدافزار تغییر نمی‌دهند. عملیات بر روی تعداد بسیاری از ثبات‌ها عوض نمی‌شود که در نتیجه محتوای برخی از ثبات‌ها تغییر نمی‌کنند؛ و بسیاری از کدهای عملیاتی در بدنه کد نیز بلاتغییر می‌مانند. در ارتباط با جریان اجرایی بدافزار، این موتورها معمولا فرآیندی مشترک را دنبال می‌کنند و فعالیت‌ها دارای وجه اشتراک زیادی هستند. این ویژگی در ساختار بدافزارهای فراریخت، انگیزه طراحی روش جدیدی برای کشف بدافزارهای فراریخت بر اساس بیشینه معیار ضریب همبستگی و کمینه معیار فاصله در این مقاله است.

یافته‌های نویسندگان حاکی از آن است که معیار مشترکی میان بسیاری از موتورهای فراریخت وجود دارد به‌گونه‌ای که در بدنه ویروس‌ها تعداد زیادی از ثبات‌ها، دستورالعمل‌ها یا کدهای عملیاتی تکرار می‌شوند. این معیار می‌تواند پایه و اساسی برای تمایز بین خانواده‌های مختلف بدافزارها باشد. اساس روش پیشنهادی، مبتنی بر تحلیل نرخ تکرار دستورات برنامه یا همان کدهای عملیاتی و ثبات‌ها است. سپس بر اساس نرخ تکرارها و بیشینه معیار ضریب همبستگی و کمینه معیار فاصله، فایل‌های سالم و خانواده‌های مختلف بدافزارهای فراریخت را طبقه‌بندی می‌کند. ترتیب انجام فعالیت‌ها در روش پیشنهادی به‌صورت شکل (۲) است.

در روش پیشنهادی ابتدا فایل مورد بررسی به کد اسمبلی برگردان می‌شود و استاندارد سازی روی آن انجام می‌گیرد که در این مرحله هر آن چیزی که در ساختار کد فایل، شمارش را با اختلال مواجه سازد حذف خواهد شد. در ریزپردازنده ۸۰۸۶، ۱۹۱ کد عملیاتی و ۲۱ ثبات در ساختار کد وجود دارند که در روش پیشنهادی به ازای هر فایل می‌بایست در مجموع این ۲۱۲ ویژگی شمارش شوند. شمارش ویژگی‌ها برای دو دسته فایل در برداری مانند $F= [f_1,f_2,\ldots, f_N]$ ذخیره می‌گردند که f_i بیانگر تعداد شمارش ویژگی i در همه فایل‌های خانواده مورد نظر می‌باشد و $N=212$ تعداد کل ویژگی‌ها است. در بخش بعد نرمال‌سازی انجام می‌گیرد. پس از شمارش کدهای عملیاتی و ثبات‌ها، چون نرخ تکرار آن‌ها در فایل‌های مختلف بسیار متفاوت است، به منظور از بین بردن این اختلاف نرمال‌سازی مطابق فرمول (۱) انجام می‌شود.

$$w_i = \frac{f_i}{\sum_{j=1}^{N} f_j} \qquad i = 1, 2, \ldots, N \qquad (۱)$$

در فرمول (۱) ، w_i حالت نرمال‌شده f_i و بیانگر احتمال وجود ویژگی i ام در ساختار کد می‌باشد که این احتمالات در بردار $W= [w_1,w_2,\ldots, w_N]$ ذخیره می‌شوند. حال با داشتن بردار چگالی احتمال ویژگی‌ها برای هر کدام از خانواده‌های فایل‌های سالم و بدافزار، فرایند پیشنهادی جهت شناسایی یک فایل ناشناس ورودی در ادامه توضیح داده خواهد شد.

فرض کنید بردار $X= [x_1,x_2,\ldots,x_N]$ بیانگر نرخ تکرار کدهای عملیاتی در فایل ورودی ناشناس باشد. قبل از اعمال روش پیشنهادی، عملیات نرمال‌سازی مطابق فرمول (۱) بر روی آن انجام می‌گیرد و خروجی با بردار $Y = [y_1,y_2,\ldots,y_N]$ نشان داده می‌شود. به عبارتی

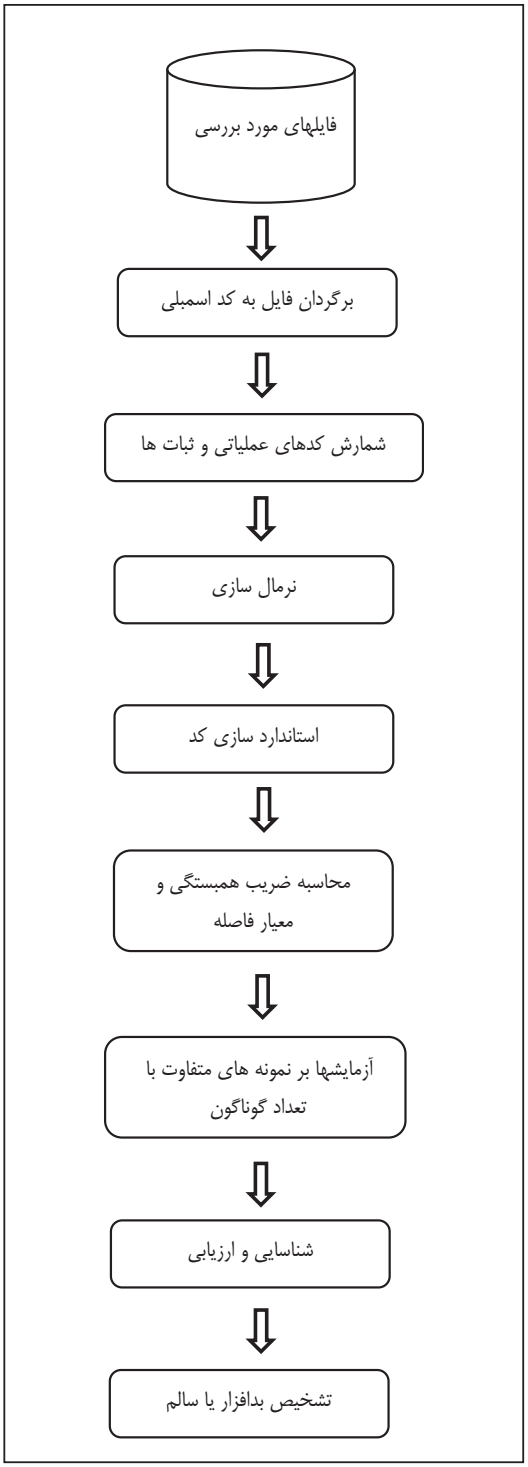
$$y_i = \frac{x_i}{\sum_{j=1}^{N} x_j} \qquad i = 1, 2, \ldots, N \qquad (۲)$$

حال برای تشخیص نوع فایل ورودی از روش‌های زیر استفاده می‌شود.

۳-۱- روش بیشینه همبستگی

در ریاضیات، معیار همبستگی یا به طور مشخص ضریب همبستگی ρ یکی از روش‌های سنجش شباهت بین دو بردار یا الگو است. ضریب همبستگی شدت رابطه و هم چنین نوع رابطه (مستقیم یا معکوس) را نشان می‌دهد .این ضریب بین ۱ تا ۱- است و در عدم وجود رابطه بین دو متغیر، برابر صفر است [۱۷].

با داشتن بردار ویژگی فایل ورودی (Y) و بردار ویژگی خانواده k ام (w^k) ضریب همبستگی از فرمول (۳) استخراج می‌گردد.



شکل (۲) : روند کلی روش پیشنهادی

$$\rho_k = \frac{\sum_{i=1}^{N} w_i^k \, y_i}{\sqrt{\sum_{i=1}^{N} (w_i^k)^2 \; \sum_{i=1}^{N} y_i^2}} \quad k = 1, 2, \dots, 5 \qquad (۳)$$

فایل ورودی عضوی از خانواده‌های فایل‌سالم و یا بدافزار k^* است که اگر به ازای آن بیشترین همبستگی به دست آید. به عبارتی

$$k^* = arg \max_k \rho_k \qquad k = 1, 2, \dots, 5 \qquad (٤)$$

۳-۲- روش کمینه فاصله

معیار تخمین فاصله (*MDE*) یک روش آماری برای اختصاص دادن یک مدل ریاضی به داده است که معمولاً آن را توزیع نمونه‌ای می‌نامند [۱۸]. در روش پیشنهادی از معیار فاصله طبق فرمول (۵) جهت سنجش میزان فاصله بردار ویژگی فایل ورودی y با هر دسته از خانواده‌های سالم و بدافزار استفاده می‌شود.

$$d_k = \sqrt{\sum_{i=1}^{N} (w_i^k - y_i)^2} \qquad k = 1, 2, \dots, 5 \qquad (۵)$$

فایل ورودی عضوی از خانواده k^* است که به ازای آن کمترین فاصله را داشته باشد. به عبارتی

$$k^* = arg \min_k d_k \qquad k = 1, 2, \dots, 5 \qquad (٦)$$

٤- ارزیابی

برای ارزیابی روش پیشنهادی، آزمایش‌های مختلفی با تعداد نمونه‌های متفاوت انجام می‌شود. فایل‌های مربوط به خانواده‌های سالم ۵۰ مورد هستند که با استفاده از ابزار *Cygwin* تولید شده‌اند و بدافزارهای فراریخت مربوط به چهار خانواده، از مرجع [۱۹] استخراج شده‌اند که در کل ۴۴۰ فایل هستند. بدافزارها از چهار خانواده فراریختی *G2, MPCGEN, MWOR, NGVCK* می‌باشند. در ابتدا همه ۴۴۰ فایل مورد بررسی، با استفاده از ابزار *IDA PRO* به کد اسمبلی برگردانده می‌شوند. پایه و اساس روش پیشنهادی در شمارش کدهای عملیاتی است که تعدادشان در ریزپردازنده ۸۰۸۶ در زبان اسمبلی ۱۹۱ مورد می‌باشد. همچنین تعداد ثبات‌ها که به صورت ساختار کد موجود هستند نیز ۲۱ مورد است. تمامی ۱۹۱ و ۲۱ مورد که در مجموع ۲۱۲ می‌شوند برای تمامی ۴۴۰ فایل مورد بررسی، در این مرحله شمارش می‌شوند. در مقایسه تحلیل‌ها مشهود است که برخی از کدهای عملیاتی فقط در برنامه‌های سالم هستند و در بدافزارها وجود ندارند. بعضی کدهای عملیاتی ممکن است در بدافزارها وجود داشته باشند، اما در برنامه‌های سالم وجود ندارند. برخی از کدهای عملیاتی نه در بدافزارها وجود دارند و نه در برنامه‌های سالم که این موارد اساس تمایز بین خانواده‌های مختلف بدافزار و سالم در روش پیشنهادی بوده است. آزمایش‌ها در حالات مختلف برای ۵۰، ۱۰۰، ۱۵۰، ۲۰۰، ۲۵۰ و ۴۴۰ فایل متشکل از فایل‌های سالم و چهار خانواده بدافزار فراریخت انجام می‌گیرد. در هر کدام از این حالات، به نسبت تعداد نمونه‌های خانواده‌های مختلف افزایش می‌یابد. تعداد دقیق نمونه‌ها در جدول (۱) نشان داده شده است.

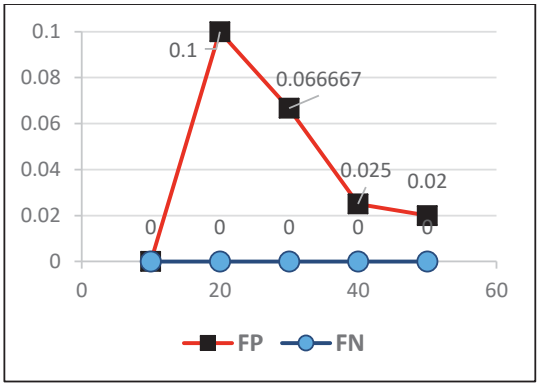
جدول (۱) : تعداد نمونه‌ها برای حالات مختلف آزمایش

فایل‌ها / تعداد	سالم	G2	MPCGEN	NGVCK	MWOR
۵۰	۱۰	۱۰	۱۰	۱۰	۱۰
۱۰۰	۲۰	۲۰	۲۰	۲۰	۲۰
۱۵۰	۳۰	۳۰	۳۰	۳۰	۳۰
۲۰۰	۴۰	۴۰	۴۰	۴۰	۴۰
۲۵۰	۵۰	۵۰	۵۰	۵۰	۵۰
۴۴۰	۴۰	۵۰	۵۰	۲۵۰	۵۰

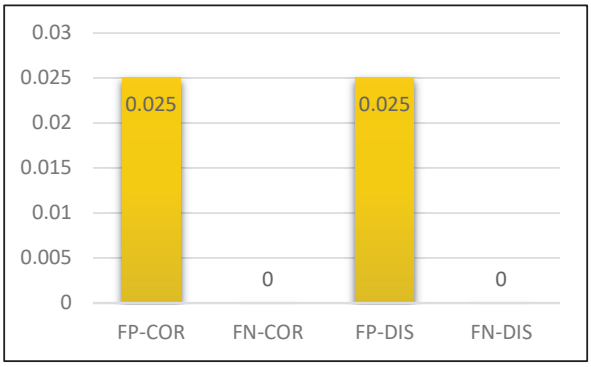
دلیل انتخاب تعداد نمونه‌های متفاوت در حالات گوناگون بررسی این مسئله بوده است که تعداد فایل‌های نمونه چه تاثیری در دقت روش پیشنهادی خواهد داشت. در ارزیابی و تحلیل نتایج، معیار *FP* نرخ مثبت کاذب و *FN* نرخ منفی کاذب می‌باشد. همچنین نمودار *ROC* جهت تخمین دقت دسته‌بندی استفاده می‌گردد که حداکثر مقدار ممکن برای آن مقدار یک و کمترین مقدار آن صفر است.

در شکل (۳) معیار *FP* و *FN* روش‌های ضریب همبستگی و معیار فاصله، در آزمایش‌های با نمونه‌های مختلف با تعداد ۵۰، ۱۰۰، ۱۵۰، ۲۰۰، ۲۵۰ و در شکل (۴) معیار *FP* و *FN* در همین دو روش، در آزمایش با ۴۴۰ نمونه نشان داده شده است. ملاحظه می‌شود که احتمال *FN* در همه نمونه‌ها صفر است. همچنین با افزایش جمعیت نمونه‌ها احتمال *FP* کاهش می‌یابد.

در جدول (۲) نیز مقایسه معیار *ROC* در روش پیشنهادی با چند مقاله معتبر نشان داده شده است. با مشاهده نتایج مشهود است که در روش پیشنهادی میزان دقت نسبت به روش‌های پیشین بیشتر بوده است و به وضوح می‌توان دید که نرخ *FP* و *FN* نیز مقدار قابل قبولی را دارد. همچنین یکی از نقاط ضعف در روش‌های پیشین این بود که با جایگزین کردن دستورات مشابه مانند *JE* و *JZ* و یا *REPE* و *REPZ* که دقیقا یک عملیات را انجام می‌دهند اما ساختار لغوی متفاوت دارند آن روش‌ها دچار شکست می‌شد. اما برای روش‌های پیشنهادی در این مقاله تمامی دستورات مشابه در شمارش‌ها یکسان در نظر گرفته شده‌اند و با جایگزین کردن آن‌ها، تاثیری در دقت شناسایی مشاهده نمی‌شود که بهبود مناسبی نسبت به روش‌های پیشین است.



شکل (۳) : معیار *FP* و *FN* در ضریب همبستگی و معیار فاصله در آزمایش حالات با تعداد ۵۰، ۱۰۰، ۱۵۰، ۲۰۰، ۲۵۰



شکل (٤) : معیار *FP* و *FN* در ضریب همبستگی و معیار فاصله در آزمایش روش پیشنهادی با ٤٤٠ نمونه

جدول (۲) : مقایسه معیار *ROC* در روش پیشنهادی با مقالات دیگر

روش‌ها	ROC
روش پیشنهادی	۱
روش مقاله [۶]	۰/۹۸۷
روش مقاله [۱۴]	۰/۹۵۷۸
روش مقاله [۲۰]	۰/۹۷۴

۵- نتیجه‌گیری و کارهای آینده

در روش پیشنهادی این مقاله با تمرکز بر شمارش و تحلیل کدهای عملیاتی و ثبات‌ها بدافزارهای فراریخت شناسایی می‌شوند. یافته‌های آزمایش‌ها نمایانگر این است که با وجود سادگی، روش ارائه شده بسیار دقیق می‌باشد. در روش پیشنهادی سربار محاسباتی بسیار کم است، ایده اساسی روش به سادگی قابل فهم و در ضدبدافزارها به آسانی قابل پیاده‌سازی است، در مقابل روش‌های جایگزینی دستورات مشابه مقاوم می‌باشد و دارای کارایی مطلوبی از نظر حافظه مصرفی و سرعت شناسایی می‌باشد. همچنین در مقایسه با روش‌های مشابه موجود دارای بهبودهایی در ابعاد مختلف است به صورتی که در آزمایش‌ها با حالات مختلف نتایج *ROC=1* و *FP=0.025* برای آن ثبت شده است که بسیار نتایج مطلوبی در زمینه شناسایی بدافزارهای فراریخت است.

برای کارهای آینده می‌توان کارایی روش را هم از نظر سرعت و هم از نظر حافظه ارتقا داد. همچنین در نظر گرفتن معیارهای موثر دیگری نظیر طراحی روش‌هایی مبتنی بر تحلیل پویا برای بررسی تغییرات در ثبات پرچم و یا در تحلیل ایستا استفاده از الگوریتم شبکه‌های عصبی مبتنی بر همین ایده نرخ تکرار کدهای عملیاتی می تواند مدنظر قرار گیرد.

مراجع

[1] Baysa, D., Low, R. M., & Stamp, M.*Structural entropy and metamorphic malware*. Journal of computer virology and hacking techniques, 9(4), 179-192, 2013.

[2] McGraw, G., Morrisett, G., "Attacking malicious code: A report to the infosec research council ", In Proceedings of IEEE Software, pp 33–44, 2000.

[3] Konstantinou, E., & Wolthusen, S. Metamorphic virus: Analysis and detection. Royal Holloway University of London, 15, 2008.

[4] Chen, L., Li, T., Abdulhayoglu, M., & Ye, Y. Intelligent malware detection based on file relation graphs. In Semantic Computing (ICSC), International Conference on (pp. 85-92). IEEE,2015.

[5] Stamp, M. *Information Security: Principles and Practice*. Wiley, New York, 2011.

[6] Canfora, G., Mercaldo, F., Visaggio, C. A., & Di Notte, P. Metamorphic malware detection using code metrics. Information Security Journal: A Global Perspective, 23(3), 57-67,2014.

[7] https://www.mcafee.com/us/resources/reports/rp-quarterly-threats-dec-2017.pdf

[8] Yanfang Y., Tao L., Qingshan J., Youyu W., "CIMDS: Adapting Postprocessing Techniques of Associative Classification for Malware Detection", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 3, pp. 298-307, 2010.

[9] Runwal, N., Low, R. M., & Stamp, M. Opcode graph similarity and metamorphic detection. Journal in Computer Virology, 8(1-2), 37-52,2012.

[10] Toderici, A. H., & Stamp, M. Chi-squared distance and metamorphic virus detection. Journal of Computer Virology and Hacking Techniques, 9(1), 1-14,2013.

[11] Al Daoud, E., Jebril, I. H., &Zaqaibeh, B. Computer virus strategies and detection methods. Int. J. Open Problems Compt. Math, 1(2), 12-20,2008.

[12] B. Anderson, D. Quist, J. Neil, C. Storlie, T. Lane. Graph-based malware detection using dynamicanalysis, J. Comput. Virol. Vol.7, No. 4, pp. 247-258,2011.

[13] Mohamed, G. A., & Ithnin, N. BSBRT: API Signature Behaviour Based Representation Technique for Improving Metamorphic Malware Detection. In International Conference of Reliable Information and Communication Technology (pp. 767-777). Springer, Cham. . 2017, April.

[14] Canfora, G., Iannaccone, A. N., &Visaggio, C. A. Static analysis for the detection of metamorphic computer viruses using repeated-instructions counting heuristics. Journal of Computer Virology and Hacking Techniques,10(1), 11-27,2014.

[15] Mehra, V., Jain, V., & Uppal, D. DaCoMM: Detection and Classification of Metamorphic Malware. Fifth International Conference on (pp. 668-673). IEEE,2015.

[16] Khammas, B. M., Monemi, A., Ismail, I., Nor, S. M., &Marsono, M. N. Metamorphic Malware Detection Based on Support Vector Machine Classification of Malware Sub-Signatures. TELKOMNIKA (Telecommunication Computing Electronics and Control), 14(3), 2016.

[17] Yates, R. D., & Goodman, D. J. Probability and stochastic processes: a friendly introduction for electrical and computer engineers (Vol. 41). Hoboken, NJ: John Wiley & Sons, 2005.

[18] Drossos, C. A., & Philippou, A. N.. A note on minimum distance estimates. Annals of the Institute of Statistical Mathematics, 32(1), 121-123, 1980.

[19] Mark Stamp Website in San Jose State University, [Online], http://cs.sjsu.edu/~stamp/viruses/

[20] Golbaghi, H., Vahidi-Asl, M., Khalilian, A., "A New Approach for Metamorphic Malware Detection by Static Analysis of Registers and Opcodes", Computing Science Journal, Vol. 4, pp. 3-15, (in Persian), 2017

زیرنویس‌ها

1 Malware
2 Malicious Malware
3 Virus
4 Worm
5 Trojan
6 Backdoor
7 Spyware
8 Ransomware
9 Rootkit
10 Metamorphic Malware
11 Obfuscation Code
12 Static Analysis
13 Dynamic Analysis
14 Signature Base Detection
15 Operation Code
16 Register
17 Correlation Coefficient
18 Minimum Distance Estimation