نشریه علمی پدافند غیرعامل

سال دهم، شماره ۱، بهار ۱۳۹۸، (پیاپی ۳۷): صص ۱۲۲-۱۰۹

تحلیل و شناسایی رفتارهای انتشاری کرم‌ها

محمدهادی علائیان[۱]، شیدا صادق‌نیا [۲]، سعید پارسا[۳]*

تاریخ دریافت: ۱۳۹۶/۰۹/۳۰

تاریخ پذیرش: ۱۳۹۷/۰۷/۲۵

چکیده

افزایش روزبه‌روز بدافزارها یکی از مهم‌ترین چالش‌های امنیت و شبکه‌های ارتباطی است. نرم‌افزارهای مخرب از لحاظ مالی و جانی به افراد و سازمان‌ها خسارت وارد می‌کنند. یکی از انواع برنامه‌های مخرب کرم‌ها هستند که از طریق ایمیل، پیام، شبکه نظیر به نظیر و اینترنت به‌صورت خودکار گسترش می‌یابند. لذا رفتارهای انتشاری موجود در کرم‌ها ما را در تشخیص کرم‌ها یاری خواهند کرد. در این راستا باید برنامه‌های سالم و مخرب در جعبه شنی اجرا گردند تا فراخوانی‌های سیستمی که تعامل برنامه با سیستم‌عامل هستند، مورد نظارت قرار گیرد. با مشاهده دنباله فراخوانی‌های سیستمی و استخراج فراخوانی‌های مربوط به انتشار می‌توان رفتار و ویژگی‌های انتشاری را به‌دست آورد. یک مجموعه از توابع سیستمی که به‌عنوان رفتار انتشاری تعیین می‌گردد را می‌توان به‌عنوان ویژگی‌های انتشاری تعریف کرد. لذا از این ویژگی‌ها در تعیین خانواده‌های کرم‌ها استفاده شده است. دقت مطلوب ۱۰۰٪ در تشخیص نشان‌دهنده این امر خواهد بود که رفتارهای انتشاری به‌درستی انتخاب‌شده‌اند. همچنین، جهت مقایسه، از الگوریتم آپریوری برای استخراج ویژگی استفاده شده است که توانسته خانواده‌های کرم‌ها را با دقت۹۶/۶۶٪ از هم متمایز کند.

کلیدواژه‌ها: تشخیص کرم، رفتارهای انتشاری، دنباله فراخوانی‌های سیستمی

۱- دانشجوی دکتری، دانشگاه علم و صنعت ایران

۲- دانشجوی کارشناسی ارشد، دانشگاه علم و صنعت ایران

۳- دانشیار، دانشگاه علم و صنعت ایران، (Parsa@iust.ac.ir) - نویسنده مسئول

۱- مقدمه

بدافزارها قطعه کدهایی هستند که بدون اجازه مالک سیستم، آن را آلوده و اقدام به کارهای ناخواسته یا خرابکارانه می‌کنند. با فراگیر شدن اینترنت و افزایش تعداد کاربران متصل به اینترنت امکان انتشار بدافزار بیشتر شده است. به‌طوری‌که طبق گزارش موسسه امنیتی مک آفی تعداد نمونه‌های شناسایی‌شده از ۴۰۰ میلیون بدافزار در اواسط سال ۲۰۱۶ به بیش از ۶۰۰ میلیون نمونه بدافزار در اواخر سال ۲۰۱۶ افزایش یافته است. از سوی دیگر تعداد نمونه‌های بدافزاری از کمتر از ۴ میلیون مورد در ابتدای سال ۲۰۱۵ به بیش از ۱۴ میلیون مورد در سه‌ماهه چهارم سال ۲۰۱۶ افزایش یافته است.

برای هر بدافزار جدید که شناخته می‌شود، باید قابلیت تخریب، بردار انتشار و تأثیر آن بر روی سیستم‌ها درک شود. این اطلاعات برای شناسایی مدل انتشار، ایجاد امضای تشخیص و روش حذف آن مهم است [۱]. روش‌های مختلف تحلیل بدافزار به تحلیلگران امنیتی اجازه می‌دهند که بدافزارها را شناسایی و با تهدیدات به‌وجودآمده از جانب آن‌ها مقابله کنند. تحلیلگران امنیتی از روش‌های ایستا و پویا برای تحلیل، تشخیص و شناسایی کرم‌ها استفاده می‌کنند. در روش مبتنی‌بر امضا، بدافزارها از روی امضاهایی که از پیش تعریف‌شده‌اند و در پایگاه داده وجود دارند، شناسایی می‌شوند. از معایب روش مبتنی‌بر امضا می‌توان به ناتوانی این روش در تشخیص بدافزارهای ناشناخته و کرم‌هایی که خود را با استفاده از روش‌هایی، از قبیل مبهم‌سازی، بسته‌بندی و رمزگذاری ایمن کرده‌اند، اشاره کرد [۳- ۲].

ازآنجایی‌که روش‌های تحلیل مبتنی بر امضا در تشخیص کرم‌های ناشناخته ناتوان بودند، تحلیلگران برای کشف روش‌های جدید مبتنی‌بر تحلیل ایستا، برای تشخیص کرم‌های ناشناخته تلاش کردند. در این راستا، تحلیل پویا که مبتنی‌بر رفتار اجرای کرم است، مطرح می‌گردد. در این روند، بدافزار بر اساس رفتارش یعنی تعاملاتی که با سیستم دارد، شناسایی می‌شود. بدین منظور کرم در یک جعبه شنی اجرا می‌شود و رفتارهای حاصل از آن تحلیل می‌گردد.

برای تشخیص کرم‌های ناشناخته از طریق فراخوانی سیستمی می‌توان مرجع‌های [۴ و ۵] را نام برد. این روش‌ها ابتدا فراخوانی‌های سیستمی را با اجرای کرم‌ها به‌دست می‌آورند. سپس هرکدام از این روش‌ها از الگوریتم‌های داده‌کاوی برای تشخیص کرم‌های ناشناخته استفاده می‌کنند. همچنین برای تحلیل کرم‌ها در حالت پویا در شبکه می‌توان مرجع‌های [۷- ۶] را نام برد. در این روش‌ها ابتدا ویژگی‌های شبکه استخراج می‌شود. سپس از الگوریتم‌های داده‌کاوی به‌منظور طبقه‌بندی داده‌ها و تشخیص کرم‌ها استفاده می‌شود. در روش‌های تحلیل مبتنی‌بر فراخوانی‌های سیستمی از فراخوانی‌های سیستمی به‌عنوان ویژگی استفاده؛ می‌گردد. اما در روش‌های تحلیل در شبکه از معیارهایی از قبیل پورت منبع، پورت مقصد، برچسب زمانی و .... به‌عنوان ویژگی استفاده می‌شود. این در حالی است که ما از رفتارهای انتشاری جهت تشخیص و شناسایی کرم‌ها استفاده می‌کنیم. ابتدا برنامه‌های سالم و مخرب در جعبه شنی به‌صورت تحلیل پویا، اجرا می‌شوند و فراخوانی‌های سیستمی مربوط به آن‌ها مورد نظارت قرار می‌گیرد. سپس با تحلیل تک‌تک این فراخوانی‌های سیستمی می‌توان رفتارهای انتشاری کرم‌ها را استخراج کرد. فراخوانی‌های مربوط به رفتارهای انتشاری به‌عنوان ویژگی در تشخیص کرم‌ها استفاده می‌شوند. در روش پیشنهادی علاوه بر تشخیص کرم‌ها می‌توان رفتار انتشار آن‌ها را شناسایی کرد. بنابراین، کاربر می‌تواند با توجه به رفتارهای انتشاری، مانع از تکثیر و انتشار کرم‌ها شود و راه‌های انتشار آن‌ها را مسدود کند.

از این‌رو، در این مقاله با اجرای کرم‌ها و فایل‌های اجرایی سالم در جعبه شنی، فراخوانی‌های سیستمی به‌صورت یک گزارش در یک فایل متنی ذخیره می‌شود. تحلیل و بررسی این فراخوانی‌های سیستمی، تمامی رفتارهای مربوط به برنامه‌های سالم و مخرب حاصل می‌گردند. سپس توابع انتشاری به‌عنوان ویژگی‌های انتشاری در تشخیص کرم‌ها استفاده می‌شود. یک کرم به‌منظور انتقال و فعال‌سازی خود بر روی سیستم‌های میزبان نیاز به شناسایی، انتقال و فعال‌سازی خود بر روی سیستم میزبان دارد. لذا توابع مربوط به شناسایی سیستم میزبان، انتقال و فعال‌سازی کرم بر روی سیستم میزبان توابع انتشاری نام‌گذاری می‌شود.

در ادامه به‌منظور تشخیص کرم‌ها، وجود هر یک از ویژگی‌ها در فایل‌های متنی حاصل از اجرای کرم‌ها و برنامه‌های سالم جست‌وجو می‌شود. سپس بانک اطلاعاتی ایجاد می‌شود. در پایان از الگوریتم جنگل تصادفی برای ایجاد مدل یادگیری و تشخیص نمونه کرم‌ها استفاده می‌شود. این ویژگی‌ها را ازآن‌جهت می‌توان مهم دانست که مربوط به رفتارهای انتشاری کرم‌ها هستند و دقت بالایی در تشخیص کرم‌ها دارند. همچنین برای مقایسه دقت ویژگی‌های استخراج‌شده، از روش آپریوری به‌منظور استخراج ویژگی استفاده می‌شود که در مرجع‌های [۹- ۸] به ترتیب برای تشخیص نشانی‌های وب فیشینگ و الگوهای خطر از بدافزارهای اندروید استفاده می‌شود. ازآنجایی‌که الگوریتم آپریوری بهترین الگوریتم شناخته‌شده در استخراج قاعده وابستگی در مقیاس بزرگ است بنابراین، می‌توان از این الگوریتم در جهت تأیید دقت ویژگی‌های انتشاری استفاده کرد. سپس دو دسته ویژگی

استخراج‌شده بر روی مجموعه کرم‌ها اعمال می‌شود. در پایان دقت هر دودسته ویژگی در تشخیص کرم‌ها مقایسه می‌شود. ویژگی‌های انتشاری در مقایسه با ویژگی‌هایی که با استفاده از الگوریتم آپریوری حاصل شده‌اند، دقت بالاتری دارد. دقت ویژگی‌های انتشاری در نمونه آماری ما ۱۰۰٪ است درحالی‌که دقت به‌دست‌آمده از ویژگی‌های آپریوری ۹۶/۶۶٪ است. همچنین روش پیشنهادی در این مقاله در مقایسه با روش‌های قبلی با دقت بالاتری در تشخیص فایل‌های مخرب عمل می‌کند به‌گونه‌ای که در مرجع [۶] نرخ تشخیص در فایل‌های ناشناخته با استفاده از الگوریتم درخت تصمیم و شبکه عصبی به ترتیب ۹۳/۲ و ۹۴/۲۷ است اما در روش پیشنهادی، دقت تشخیص و طبقه‌بندی با استفاده از الگوریتم درخت تصمیم و شبکه عصبی به ترتیب ۱۰۰ و ۹۳/۳۳ است.

بنابراین، با استفاده از ویژگی‌های انتشاری می‌توان فایل‌های ناشناخته را با دقت بالاتری نسبت به روش‌های قبلی شناخت. همچنین رفتارها و راه‌های انتشار، فایل‌های ناشناخته حاصل می‌شود و تحلیلگر به‌راحتی می‌تواند این راه‌های انتشاری را مسدود کند.

در ادامه این مقاله، در بخش دوم، روش انتشار کرم‌ها توضیح داده‌شده است. در بخش سوم مروری بر مطالعات پیشین در زمینه تحلیل، تشخیص و شناسایی کرم‌ها ارائه شده است. در بخش چهارم، روش پیشنهادی توضیح داده شده است و در بخش پنجم، نتایج آزمایش‌ها درج شده است.

۲-روش‌های انتشار

نکته مهم پس از دانستن درباره کرم‌ها، نحوه شناسایی و جلوگیری از انتشار آن‌ها است. کرم‌ها در اصطلاح به نرم‌افزارهای مخربی گفته می‌شود که با اهداف مختلفی ازجمله جمع‌آوری اطلاعات حساس، دسترسی به سیستم‌های رایانه‌ای خصوصی و در برخی موارد تخریب سیستم‌ها در شکل‌های گوناگون مانند اسکریپت، کد، محتوای فعال و ... طراحی‌شده و با کمک عوامل انسانی یا به‌صورت خودکار و به شیوه‌های خاص و رسانه‌های چندگانه در بین رایانه‌ها منتشر می‌شوند. به روش‌های انتقال بدافزارها از یک سیستم به سیستم‌های دیگر انتشار بدافزار گفته می‌شود و بر روی آن تأثیر می‌گذارد. از رایج‌ترین راه‌های انتشار و انتقال کرم‌ها، موارد زیر را می‌توان نام برد:

۲-۱- انتشار از طریق ایمیل

نفوذ گران به‌منظور ترغیب افراد به باز کردن ضمیمه ایمیل، از اسامی و عنوان‌های جذاب که حس کنجکاوی افراد را برمی‌انگیزد، استفاده می‌کنند. هنگامی‌که فرد مخرب به یک سیستم دسترسی داشته باشد درواقع به فهرست آدرس ایمیل‌های استفاده‌شده توسط آن سیستم نیز دسترسی خواهد داشت. فایل‌هایی با پسوند pst دارای فهرستی از ایمیل‌های افراد است که توسط نرم‌افزار اوت‌لوک[1] مورد استفاده قرار می‌گیرد. بنابراین، می‌تواند فایل‌های مخرب خود را به این آدرس ایمیل‌ها ارسال کند تا موجب گسترش بدافزار گردد. ازجمله این نوع از کرم‌ها Loveletter، Melica، Happy99، AnnaKournikova و My life را می‌توان نام برد.

۲-۲- انتشار از طریق پیام

کرم‌ها به دو صورت انتقال فایل و انتقال لینک از طریق پیام انتشار می‌یابند. در روش اول هکرها با ضمیمه قرار دادن کرم‌ها به پیام‌های دارای محتوای مخرب و جلب‌توجه کاربر جهت اجرای کرم و درنهایت منجر به آلوده شدن سیستم می‌شود. در روش دوم سیستم آلوده یک لینک که حاوی کرم است را برای یکی از کلاینت‌ها ارسال خواهد کرد. کلاینت بر روی لینک کلیک خواهد کرد و به سایت مربوط به کرم ارجاع داده می‌شود. سپس کرم به‌طور خودکار دانلود خواهد شد و بر روی سیستم میزبان نصب می‌گردد. ازجمله این کرم‌ها Choke، JSMenger، Kelvir/Bropia، Serflog را می‌توان نام برد.

انتشار کرم‌های ایمیل و نمونه پیام دقیقاً شبیه است. تنها تفاوت آن‌ها در چهار مورد زیر است:

- ✓ محتوای پیغام: محتوای پیام کرم‌های ایمیل با کرم‌های نمونه پیام متفاوت است.
- ✓ روش‌های ارسالی: کرم‌های ایمیل از ایمیل‌ها، کرم‌های نمونه پیام از مسنجرها به‌عنوان ابزارهایی برای ارسال استفاده می‌کنند.
- ✓ تعداد میزبان‌ها: تعداد میزبان‌ها در کرم‌های نمونه پیام و ایمیل متفاوت است.
- ✓ نحوه دسترسی: نحوه دسترسی از راه دور بر روی سیستم میزبان.

در روش مهندسی اجتماعی هکرها به‌جای استفاده از روش‌های معمول و مستقیم نفوذ، جهت جمع‌آوری اطلاعات و عبور از دیواره آتش برای دسترسی به سیستم‌ها، از طریق مسیرهای انسانی و فریب دادن کاربران به جمع‌آوری اطلاعات درون سیستم می‌پردازند [۲].

پیوست‌های ایمیل جهت ارسال کدهای مخرب به سیستم قربانی استفاده می‌شود. از مهندسی اجتماعی استفاده کرد که می‌تواند به‌طور خودکار چیزی مثل یک نرم‌افزار واسط بین صفحه‌کلید و کامپیوتر[2] به‌منظور گرفتن گذرواژه‌ها را اجرا کند.

1- outlook
2- Keylogger

ویروس‌ها، تروجان‌ها و کرم‌ها را می‌توان هوشمندانه در ایمیل‌های دست‌کاری‌شده قرار داد تا قربانی را با باز کردن آن‌ها وسوسه کند. این یک مثال از ایمیلی است که تلاش می‌کند تا دریافت‌کننده را برای باز کردن یک فایل پیوست ناامن متقاعد کند. به نظر می‌رسد که این ایمیل از طرف یک دوست یا شخص آشنا ارسال‌شده است و کاربر به باز کردن آن رغبت دارد مانند یک شرکت بانکی یا کارمند همکار. این ایمیل ممکن است حاوی یک لینک باشد که به وب‌سایت جعلی ارجاع داده می‌شود [۴].

هکرها به‌منظور ترغیب افراد به باز کردن ضمیمه ایمیل، از اسامی و عناوین جذاب که حس کنجکاوی افراد را برمی‌انگیزد، استفاده می‌کنند. هنگامی‌که هکر به یک سیستم دسترسی داشته باشد درواقع به ایمیل‌های آن سیستم نیز دسترسی خواهد داشت بنابراین، می‌تواند ایمیل‌های ضمیمه‌شده را از طریق ایمیل سیستمی که به آن دسترسی دارد، برای همه مخاطبان قربانی ارسال کند. این ایمیل می‌تواند حاوی لینک، پیغام و دانلودها باشد.

۲-۳- انتشار از طریق شبکه نظیر به نظیر

در این روش کرم‌ها از طریق پوشه‌های[1] مشترک در شبکه‌های نظیر به نظیر[2] انتشار می‌یابند. به‌عنوان‌مثال، کرم‌های نظیر به نظیر در این نوع انتشار از برنامه پوشه‌های مشترک LimeWire برای دانلود کرم‌ها بر روی سیستم میزبان استفاده می‌کنند. LimeWire سریع‌ترین نرم‌افزار به اشتراک‌گذاری فایل‌ها در دنیا است. به‌وسیله این نرم‌افزار می‌توان به حجم عظیمی از فایل‌های tiff، Avi ،Jpg ،Mp3 دسترسی داشت، فایل‌های موردنظر خود را جستجو کرده و یا فایل‌های خود را با دیگر کاربران به اشتراک گذاشت. از ویژگی‌های این نرم‌افزار می‌توان به‌سادگی، قابل‌اجرا بر روی سیستم‌عامل‌های مختلف، پشتیبانی از زبان‌های مختلف و ... اشاره کرد. کرم‌های نظیر به نظیر به‌منظور گسترش از روش‌های مهندسی اجتماعی برای نام‌گذاری استفاده می‌کنند و کاربران را وادار به دانلود و راه‌اندازی فایل‌ها می‌کنند.

۲-۴- انتشار از طریق اینترنت

بعضی از کرم‌ها می‌توانند خود را بدون دخالت عامل انسانی و از طریق شبکه تکثیر و منتشر کنند. کرم‌های اینترنتی اغلب حمله‌های مخربی را در مقابل شبکه‌های کامپیوتری انجام می‌دهند. مراحل حیات هر کرم اینترنتی شامل یافتن هدف، انتقال، فعال‌سازی و آلوده‌سازی است.

در مرحله هدف‌یابی، قربانی انتخاب می‌شود، یکی از ساده‌ترین روش‌ها در این مرحله پویش کورکورانه آدرس‌های پروتکل اینترنت[3] است. در مرحله انتقال یک نسخه از کرم به قربانی منتقل می‌شود. در مراحل فعال‌سازی و آلوده‌سازی کرم روی ماشین قربانی اجراشده و آن را آلوده می‌کند.

۳- کارهای مرتبط

به‌منظور افزایش امنیت شبکه‌ها در چند سال اخیر، تشخیص و طبقه‌بندی بدافزارها اهمیت خاصی پیداکرده است. کارهای مرتبط در زمینه تشخیص بدافزارها را می‌توان به دودسته مبتنی بر رفتار و مبتنی بر شبکه دسته‌بندی کرد.

۳-۱- تشخیص مبتنی بر شبکه

بارهام[4] و همکارانش در [۶] با ترکیب روش‌های طبقه‌بندی با استفاده از درخت تصمیم[5] و شبکه عصبی[6] برای تشخیص کرم‌ها استفاده کردند. در این روش از ویژگی‌هایی همچون پورت منبع، پورت مقصد، برچسب زمانی و کد کلید مجازی به‌منظور تشخیص بدافزارها بهره‌مند شدند. روش بارهام و همکارانش با دقت بالایی کرم‌های ناشناخته را تشخیص می‌دهد. به‌گونه‌ای که دقت تشخیص کرم‌های ناشناخته بر اساس درخت تصمیم و شبکه عصبی به ترتیب ۹۳/۲٪ و ۹۴/۲۷٪ است درحالی‌که دقت تشخیص کرم‌های ناشناخته با ترکیب این دو روش ۹۷/۷۴٪ است. ولی اگر کرم با اینترنت تعامل کمی داشته باشد در این صورت این روش نمی‌تواند کار تحلیل را به‌خوبی انجام دهد.

روش دیگری که در مرجع [۷] برای تحلیل، تشخیص و شناسایی کرم اینترنت استفاده‌شده است، روش طبقه‌بندی مبتنی بر شنود بر روی پورت و اعمال حملات محروم‌سازی از سرویس[7] است. برای هر رکورد ۱۳ ویژگی وجود دارد که در یک ثانیه از کل بسته جست‌وجو می‌شود. در مرجع [۷] هر آدرس پروتکل اینترنت منبع با در نظر گرفتن فاصله زمانی یک ثانیه، یک رکورد است. ازجمله این ویژگی‌ها، آدرس پروتکل اینترنت منبع، آدرس پروتکل اینترنت مقصد، تعدادی از بسته‌های پروتکل کنترل انتقال[8] و پروتکل کنترل پیام‌های اینترنتی[9] و مجموعه‌ای از پورت مقصد، مجموعه‌ای از پورت منبع و ... را می‌توان نام برد. پس از استخراج ویژگی از سه الگوریتم درخت تصمیم، شبکه عصبی، جنگل تصادفی[10] به‌منظور طبقه‌بندی داده‌ها استفاده می‌شود. در این پژوهش الگوریتم جنگل تصادفی با نرخ تشخیص ۹۹/۶ از دو الگوریتم درخت تصمیم و شبکه عصبی با دقت ۹۹/۴ و ۹۷/۸

1- Folder
2- P2P networks

3- IP
4- Barhoom
5- Decision Tree
6- Artificial Neural Networks
7- DOS
8- TCP
9- ICMP
10- RandomForest

بهتر عمل می‌کند. البته باید گفت این روش تعداد محدودی از کرم‌هایی که با شبکه در تعامل هستند خصوصاً کرم‌های اینترنت را می‌تواند شناسایی کند و سایر کرم‌ها می‌توانند به کار مخرب خود ادامه دهند.

۳-۲- تشخیص مبتنی بر فراخوانی سیستمی

در مرجع [۵] از الگوریتم DTW[1] برای تشخیص بدافزارها استفاده کردند. با بررسی رفتار کرم‌ها، فایل‌های مخرب و سالم در یک جعبه شنی اجرا می‌شوند و فراخوانی‌های سیستمی حاصل از آن‌ها نیز استخراج می‌گردد. مجموعه‌ای از فراخوانی‌های سیستمی یک برنامه که کار هدفمندی را انجام می‌دهد را می‌توان رفتارهای برنامه مربوطه دانست. سپس تمایز توابع سیستمی فراخوانی شده توسط فایل‌های مخرب از سالم، رفتارهای مخرب را تشکیل می‌دهد. این روش با بهره‌گیری از ردیابی فراخوانی سیستمی و محاسبه ماتریس فاصله با استفاده از الگوریتم DTW برای شکل‌دهی و تشکیل گروه بدافزارها استفاده می‌کند. مزیت اصلی این روش تشخیص سریع‌تر بدافزارها و عدم نیاز به بررسی تک‌تک بدافزارها است. این روش بیش‌تر برای تشخیص نمونه‌هایی از بدافزارهای ناشناخته استفاده می‌کند.

همچنین، در مرجع [۴] یک روش تشخیص بدافزار به‌منظور تحلیل بدافزارهای ناشناخته ارائه شده است. در این روش ابتدا باید با استفاده از داده‌های خام، بردار ویژگی برای هر نمونه تولید شود. در این حالت از مدل bag of words به‌منظور استخراج ویژگی استفاده می‌شود. در این مدل به‌جای تمام کلمات، تعدادی از آن‌ها انتخاب می‌شوند. برای مثال، کلمه‌ای مانند is نمی‌تواند نشان‌دهنده نوع یک متن باشد پس می‌توان آن‌ها را از بردار ویژگی حذف کرد. در این مطالعه به‌طور مشابه از مدل bag-of-n استفاده می‌شود. این روش ردیابی فراخوانی سیستمی را به‌صورت بردار فرکانس از n-gram نمایش می‌دهد. سپس به‌منظور کاهش ابعاد ویژگی‌ها از الگوریتم محاسبه فرکانس واژه نرمال شده فرکانس سند معکوس[2] استفاده می‌شود. این الگوریتم به‌منظور بررسی اهمیت یک کلمه در متن مورداستفاده قرار می‌گیرد. منظور از اهمیت در این قسمت تعداد دفعاتی است که یک کلمه در متن تکرار شده است. این الگوریتم می‌تواند برای فیلتر کردن در زمینه‌های مختلف از جمله خلاصه متن و طبقه‌بندی استفاده شود. در این مطالعه از الگوریتم محاسبه فرکانس واژه نرمال شده - فرکانس سند معکوس به‌منظور کاهش ابعاد ویژگی استفاده می‌شود. در پایان از الگوریتم‌های تشخیص مبتنی بر امضا، آزمون نسب معادله ورودی چندضلعی[3]، رگرسیون لجستیک[4] به‌منظور تشخیص بدافزارها استفاده می‌شود.

پالاهان[5] و همکارانش [۱۰] از روش تحلیل پویا به‌منظور تشخیص بدافزار استفاده کردند. آن‌ها از اطلاعات آماری توابع سیستمی فراخوانی شده توسط برنامه‌های سالم و مخرب و ایجاد گراف‌های فراخوانی سیستمی جهت تمایز بین فایل‌های سالم و مخرب استفاده کردند. رفتارهای برنامه به‌صورت گراف فراخوانی سیستمی نمایش داده می‌شود. سپس گراف به یک بردار ویژگی برای تولید یک ویژگی برای توالی فراخوانی سیستمی تبدیل می‌شود. تعداد دفعاتی که هر دو فراخوانی سیستمی متوالی فراخوانی می‌شوند، ویژگی‌های این روش هستند. اگر این مقدار مثبت باشد نشان‌دهنده این است که فایل مربوطه بدافزار خواهد بود اما اگر این مقدار منفی باشد نشان‌دهنده این است که فایل سالم است. نقطه‌ضعف عمده روش‌های مطرح‌شده در مرجع [۱۰][۴] عدم توجه به ویژگی‌های معنایی رفتار برنامه است.

۴- روش پیشنهادی

در این روش کرم‌ها بر اساس رفتار انتشار آن‌ها تشخیص داده می‌شوند. برای استخراج رفتار انتشار کرم‌ها، پس از اجرای فایل‌های سالم و مخرب در جعبه شنی یک فایل متنی که حاوی فراخوانی‌های سیستمی مربوط به برنامه است، حاصل می‌شود. با تحلیل این فراخوانی‌های سیستمی به‌دست‌آمده و بررسی ارتباط بین هدف بدافزار و کاربری توابع، رفتارهای انتشار کرم‌ها تعیین می‌شوند. بنابراین، می‌توان فراخوانی‌های سیستمی که مربوط به انتشار کرم‌ها هستند را استخراج کرد و از آن‌ها به‌عنوان ویژگی‌های انتشاری بدافزارها استفاده کرد. سپس پایگاه داده مربوطه با توجه به این ویژگی‌های استخراج‌شده و جست‌وجوی آن‌ها در نمونه‌های مختلف ایجاد خواهد شد و در پایان از الگوریتم‌های مناسب یادگیری ماشین جهت طبقه‌بندی داده‌ها و تشخیص کرم‌ها استفاده خواهد شد.

در این مقاله کرم‌های مخرب نرم‌افزاری بر اساس دودسته ویژگی تشخیص داده می‌شوند. دسته اول ویژگی‌هایی هستند که بر اساس رفتارهای انتشاری آن‌ها استخراج شده‌اند. دسته دوم ویژگی‌هایی هستند که با استفاده الگوریتم آپریوری[6] از فایل‌های متنی حاصل از اجرای برنامه‌های سالم و مخرب، که شامل توالی از فراخوانی‌های سیستمی است، استخراج شده‌اند. درواقع دسته دوم از ویژگی‌ها به‌منظور نشان دادن دقت بالای ویژگی‌های انتشاری استفاده می‌شود.

این روش شامل ۶ گام اصلی است:

1- Dynamic time warping
2- TF-IDF
3- LLRT
4- LR
5- Palahan
6- Appriori

۱. دسته‌بندی کرم‌ها بر اساس عملکرد آن‌ها
۲. استخراج الگوی رفتاری کرم‌ها از فراخوانی سیستمی
۳. استخراج ویژگی
۴. ایجاد بانک اطلاعاتی
۵. استفاده از الگوریتم‌های داده‌کاوی در طبقه‌بندی کرم‌ها
۶. آزمون مدل یادگیری شده و تشخیص فایل ناشناخته

در ادامه به تشریح هر یک از این مراحل خواهیم پرداخت.

۴-۱- دسته‌بندی کرم‌ها بر اساس عملکرد آن‌ها

جهت اعمال الگوریتم به خانواده‌ای از کرم‌ها نیاز خواهد بود که ما نمونه‌ها را از سایت vx-archiv.at دانلود کنیم. به این دلیل که اساساً، هر خانواده، آسیب‌پذیری‌های متفاوتی را مورد هدف قرار می‌دهد و گستره عملکردی مختلفی نسبت به مابقی خانواده‌ها دارد. به‌این‌ترتیب می‌توان با توجه به عملکردی که کرم‌ها از خود نشان داده‌اند آن‌ها را دسته‌بندی کرد. عمل دسته‌بندی می‌تواند به دو صورت دستی و خودکار صورت گیرد. در این مقاله دسته‌بندی کرم‌ها به‌صورت دستی و بر اساس گزارش‌های ارائه‌شده در مطالعات پیشین در مرجع [۱۱] انجام شده است. از آنجا که این گزارش‌ها توسط محققان امنیتی و بر اساس مشاهده عملکرد آن کرم‌ها در حین اجرا و در یک محیط واقعی صورت می‌گیرد، بسیار دقیق‌تر از روش‌های خودکار خوشه‌بندی یا دسته‌بندی است. بر اساس این دسته‌بندی کرم‌ها به چهار دسته کرم‌های ایمیل، نمونه پیام، شبکه نظیر به نظیر و اینترنت تقسیم می‌شوند.

۴-۲- استخراج الگوی رفتاری کرم‌ها از فراخوانی سیستمی

نمایش رفتار یک برنامه بر اساس تعاملاتی که آن برنامه با سیستم‌عامل از طریق فراخوانی‌های سیستمی انجام می‌دهد، توانایی قدرتمندی را در کشف بدافزار مبتنی بر رفتار ایجاد کرده است. درروش پیشنهادی برای استخراج رفتار انتشار کرم‌ها از یک جعبه شنی استفاده می‌شود. در این مرحله چندین نمونه متفاوت از هر خانواده از کرم‌ها در این جعبه شنی اجرا می‌شوند. سپس فراخوانی‌های سیستمی که در قالب یک فایل متنی از جعبه شنی به دست می‌آید، به‌منظور استخراج رفتار انتشار کرم تحلیل می‌شود. با تحلیل و استخراج عملکرد هریک از این فراخوانی‌های سیستمی بررسی ارتباط بین آن‌ها، رفتار انتشار کرم‌ها استخراج می‌شود. فراخوانی‌های سیستمی مربوط به انتشار کرم‌ها و فراخوانی‌های مربوط به اجرای برنامه‌های سالم که در کرم‌ها وجود ندارد، در جدول (۱) نشان داده شده است.

۴-۳- استخراج ویژگی

استخراج ویژگی فرآیندی است که در آن با بررسی و انجام عملیات بر روی داده‌ها، ویژگی‌های بارز مشخص می‌شود. هدف استخراج ویژگی این است که داده‌های خام برای پردازش‌های آماری آماده شوند. این مسئله در بسیاری از کاربردها مانند تشخیص و طبقه‌بندی بدافزارها اهمیت به‌سزایی دارد، اگر ویژگی‌ها به‌درستی انتخاب شوند دقت تشخیص و طبقه‌بندی بدافزارها بیشتر خواهد بود. در این مرحله سه دسته ویژگی استخراج می‌شود. دسته اول ویژگی‌های انتشاری هستند که با توجه به رفتارهای انتشاری و به‌صورت دستی استخراج می‌شوند. دسته دوم ویژگی‌های هستند که از تمامی فراخوانی‌های سیستمی و با استفاده از قوانین به‌دست‌آمده از الگوریتم آپریوری استخراج می‌شوند. دسته سوم ویژگی‌هایی هستند که با تلفیق دودسته ویژگی قبلی به‌دست می‌آیند. ویژگی‌های دسته دوم و سوم، جهت، تائید در تشخیص رفتار انتشاری، دقت در تشخیص کرم‌ها در نظر گرفته‌شده است.

۴-۳-۱- ویژگی‌های انتشاری:

در این مرحله ابتدا تمامی فراخوانی‌های سیستمی حاصل از اجرای کرم‌ها و برنامه‌های سالم بررسی می‌شود. فراخوانی‌های سیستمی مربوط به یک خانواده از کرم‌ها یکسان هستند. هر یک از این توابع سیستمی، عملی مختص به خود را انجام می‌دهند. عملکرد مربوط به هر یک از این توابع با جست‌وجو در اینترنت استخراج می‌شود. سپس با توجه به‌ترتیب توابع و بررسی این توابع متوالی توسط فرد تحلیلگر، رفتارهای مربوط به هر خانواده از کرم‌ها استخراج می‌شود. سپس توابعی که مربوط به انتشار کرم‌ها هستند و در خانواده‌های مختلف از کرم‌ها متفاوت‌اند، به‌عنوان ویژگی‌های انتشاری استخراج می‌شوند. این ویژگی‌ها فراخوانی‌های سیستمی هستند که رفتار انتشار کرم‌ها را نشان می‌دهند. ما طبق بررسی‌هایی که انجام داده‌ایم و تحلیل کرم‌های مخرب، ویژگی‌های جدول (۲) را به‌عنوان ویژگی‌های ضروری جهت تمایز بین خانواده کرم‌ها و فایل‌های سالم در نظر گرفتیم. ویژگی‌های استخراج‌شده و عملکرد مربوط به هر یک از آن‌ها در جدول (۲) نشان داده شده است. به‌منظور استخراج ویژگی، در این مرحله چندین نمونه از فایل‌های سالم جهت مقایسه با کرم‌ها و به دست آوردن تمایز بین آن‌ها در جعبه شنی مربوطه اجرا شد. وجود یا عدم وجود این ویژگی‌ها بر دقت تشخیص کرم‌ها تأثیر دارند.

اگرچه همه فراخوانی‌های مربوط به کرم‌های موجود در جدول (۱) به‌منظور انتشار کرم‌ها لازم هستند ولی بعضی از آن‌ها به‌شدت در فراخوانی‌های سالم نیز مشاهده شده‌اند و ازآنجایی‌که میزان دقت در تشخیص کرم‌ها کاهش می‌دهد. لذا ما آن‌ها را حذف کردیم. فراخوانی‌های به‌دست‌آمده از جدول (۱) که با دقت بالاتری در تشخیص کرم‌ها عمل می‌کنند در جدول (۲) به‌عنوان ویژگی‌های انتشاری نشان داده شده‌اند.

جدول (۱): فراخوانی‌های انتشاری کرم‌ها و فراخوانی‌های برنامه سالم

نام تابع	نام خانواده	نام تابع	نام خانواده
ntvdm.exe	ایمیل	Sechost.dll	ایمیل، پیام، اینترنت، نظیر به نظیر، برنامه سالم
Createfile	ایمیل، پیام، نظیر به نظیر، اینترنت، برنامه سالم	Psapi.dll	نظیر به نظیر
conhost.exe	ایمیل	shell32	ایمیل، اینترنت، نظیر به نظیر، برنامه سالم
Error Message Instrument	ایمیل، پیام، اینترنت، نظیر به نظیر، برنامه سالم	Active computer name	ایمیل، پیام، اینترنت، نظیر به نظیر، برنامه سالم
True type font	ایمیل	Send message	ایمیل، پیام، نظیر به نظیر، برنامه سالم
NtRequestWaitReplyPort	ایمیل، پیام، اینترنت، نظیر به نظیر، برنامه سالم	Read file	ایمیل، پیام، نظیر به نظیر، اینترنت، برنامه سالم
WindowsError Reporting\\WMR	ایمیل، پیام، اینترنت، نظیر به نظیر، برنامه سالم	Telemetry client	ایمیل، پیام، برنامه سالم
Gdi32.dll	ایمیل، پیام، نظیر به نظیر، اینترنت	Sqm	ایمیل، پیام، اینترنت، برنامه سالم
User32.dll	ایمیل، پیام، نظیر به نظیر، اینترنت	Load library (نام فایل نصبی کرم)	پیام، نظیر به نظیر، اینترنت
Mscvrt.dll	ایمیل، اینترنت، پیام، نظیر به نظیر، برنامه سالم	Terminal server	پیام، نظیر به نظیر، اینترنت
Msctf.dll	ایمیل، پیام، نظیر به نظیر، اینترنت، برنامه سالم	TSUserEnabled	پیام، نظیر به نظیر، اینترنت
rpcrt4.dll	ایمیل، پیام، اینترنت، نظیر به نظیر، برنامه سالم	Sysmain.sdb	ایمیل، پیام، نظیر به نظیر، اینترنت، برنامه سالم
MAXIMUM_ALLOWED	ایمیل، پیام، نظیر به نظیر، اینترنت، برنامه سالم	Full Screen	ایمیل
cryptbase.dll	ایمیل، پیام، نظیر به نظیر، اینترنت، برنامه سالم	Msvbm60.dll	پیام
bcryptprimitives.dll	ایمیل، پیام، نظیر به نظیر، اینترنت، برنامه سالم	Disable8And16 BitMitigation	اینترنت، پیام، نظیر به نظیر، ایمیل
FipsAlgorithmPolicy	ایمیل، پیام، نظیر به نظیر، اینترنت، برنامه سالم	Image file extention	پیام
Side by side	ایمیل، پیام، نظیر به نظیر، اینترنت، برنامه سالم	ws2_3	پیام
Advapi	ایمیل، پیام، نظیر به نظیر، اینترنت،	comdlg32	برنامه سالم

	برنامه سالم		
Setupap	برنامه سالم	SHCore	ایمیل، برنامه سالم
cfgmgr32	برنامه سالم	Profapi	برنامه سالم
True Type Font	ایمیل	Monitors	پیام
Host = gethostbyname ("23.223.16.227);	پیام، اینترنت	Netapi32.dll	ایمیل
Netutils.dll	ایمیل	tiptsf.dll	ایمیل، برنامه سالم
Create remote thread	ایمیل	Open thread	ایمیل، نظیر به نظیر
Write file	ایمیل، پیام، نظیر به نظیر، برنامه سالم	Exite thread	ایمیل، پیام، نظیر به نظیر، اینترنت
Exit process	ایمیل، پیام، نظیر به نظیر، اینترنت		

جدول (۲): ویژگی‌های انتشاری استخراج‌شده

نام ویژگی	عملکرد	نام ویژگی	عملکرد
Ntvdm	کرم ایمیل قادر به‌جای دادن خود در درایو C سیستم نیست به همین دلیل برای مخفی سازی و جای‌گذاری خود در این درایو از بارگذاری کردن ntvdm.exe در حافظه استفاده می‌کند.	Image file extention	اگر بخواهیم عملیات اجراشدن دیباگر بعد از اجرای یک فایل خاص به‌صورت خودکار اجرا شود. از این ویژگی در رجیستری ویندوز استفاده می‌شود. این روش توسط هکر برای اجرا کردن بدافزار پس از اجرای یک برنامه خاص استفاده می‌شود.
Terminal	بعضی از کرم‌ها به‌منظور دسترسی از راه دور بر روی سیستم میزبان از این رجیستری استفاده می‌کند.	Monitors	به‌منظور جست‌وجوی میزبان‌های آسیب‌پذیر از این رجیستری استفاده می‌شود
TSUser Enabled	با انتساب این مقدار به رجیستری، عملیات دسترسی از راه دور فعال می‌شود.	Msvbvm60	برای چک کردن فولدرهای سیستم از این کتابخانه استفاده می‌شود.
Disable8And 16BitMitigation	این تابع رمز یک سند محافظت‌شده را کرک می‌کند.	ws2_32	طیف گسترده‌ای از کنترل‌های استاندارد ویندوز مانند ذخیره و تبادل فایل را فراهم می‌کند.
CreateRemoteThread	فرآیند ریموت بر روی سیستم میزبان ایجاد می‌شود.	netapi32	به‌منظور مدیریت شبکه ایجاد می‌شود.
SHCore	قسمتی از سیستم‌عامل ویندوز است بنابراین، ممکن است از طریق نصب ویندوز به سیستم منتقل‌شده باشد.	comdlg32	نسخه‌های در دسترس از کتابخانه در دسترس رایج را لیست می‌کند و توصیف می‌کند که برنامه شما از چه نسخه‌ای استفاده می‌کند.
Setupapi	تابع راه‌اندازی مجموعه‌ای از توابع که برای نصب یک برنامه لازم است را فراخوانی می‌کند.	cfgmgr32	این کتابخانه در درایو C قرار دارد و می‌تواند به برنامه در حال اجرا تزریق شود و رفتار آن برنامه را تغییر دهد. برخی از بدافزارها خود را به‌عنوان cfgmgr32 مخفی می‌کنند.
True Type Font	فونت را برای command promat مشخص می‌کند.	Tiptsf	لودشدن این کتابخانه در حافظه باعث فعال‌سازی میزبان نسبت به پاسخگویی به درخواست سیستم اولیه می‌شود.
Profapi	یک جز نرم‌افزاری از اجزای سیستم‌عامل ویندوز است.		

۴-۳-۲- ویژگی‌های به‌دست‌آمده از قوانین آپریوری

الگوریتم آپریوری زیرمجموعه‌هایی از آیتم‌ها را که بین C مجموعه آیتم مشترک است را پیدا می‌کند. بنابراین، از آیتم‌های موجود در قوانین می‌توان به‌عنوان ویژگی‌های تشخیص کرم‌ها استفاده کرد. در ادامه به‌منظور نشان دادن دقت بالای ویژگی‌های انتشاری در تشخیص کرم‌ها از ویژگی‌های آپریوری استفاده می‌شود. بدین منظور ابتدا فایل‌های متنی که از اجرای کرم‌ها و فایل‌های سالم حاصل شد، بررسی می‌شود. سپس تمامی فراخوانی‌های سیستمی مربوط به کرم‌ها و فایل‌های سالم استخراج می‌شود. با بررسی وجود این فراخوانی‌ها در تمام فایل‌های متنی پایگاه داده ایجاد می‌شود. پایگاه داده حاصل‌شده به ابزار وکا داده می‌شود. از آنجایی که حجم داده‌ها خیلی زیاد است.

لذا الگوریتم آپریوری نمی‌تواند به‌راحتی قوانین مربوطه از آن را استخراج کند. از الگوریتم انتخاب ویژگی در قسمت پیش‌پردازش داده به‌منظور فیلتر کردن استفاده می‌شود. سپس داده‌های حاصل‌شده به الگوریتم آپریوری داده می‌شود. ویژگی‌هایی که از این قسمت حاصل می‌شود در جدول (۵) نشان داده شده است.

۴-۳-۳- تلفیق ویژگی‌های انتشاری و ویژگی‌های به‌دست آمده از الگوریتم آپریوری

در این مرحله ویژگی‌های انتشاری که به‌صورت دستی استخراج شده‌اند به همراه ویژگی‌های انتشاری به‌دست‌آمده از الگوریتم آپریوری با یکدیگر در نظر گرفته می‌شوند. ویژگی‌های به‌دست‌آمده از تلفیق این دو دسته ویژگی در جدول (۶) نشان داده شده است.

جدول (۳): ویژگی‌ها استخراج‌شده با استفاده از قوانین به‌دست‌آمده از آپریوری

ویژگی	عملکرد
Ntvdm	کرم ایمیل قادر به‌جای دادن خود در درایو c سیستم نیست به همین دلیل برای مخفی سازی و جای گذاری خود در این درایو از بارگذاری کردن ntvdm.exe در حافظه استفاده می‌کند.
Monitors	به‌منظور جست‌وجوی میزبان‌های آسیب‌پذیر از این رجیستری استفاده می‌شود
Msvbvm	برای چک کردن فولدرهای سیستم از این کتابخانه استفاده می‌شود.
SHCore	قسمتی از سیستم‌عامل ویندوز است بنابراین ممکن است از طریق نصب ویندوز به سیستم منتقل‌شده باشد.
True Type Font	فونت را برای command promat مشخص می‌کند
Full Screen	تنظیمات صفحه مربوط به پیام کرم ایمیل استفاده می‌شود.
Tiptsf	از طریق این کتابخانه میزبان برای کلیک بر روی پیغام فعال می‌شود.
Send Message	با استفاده از این تابع پیغامی برای سیستم هدف و شبکه فرستاده می‌شود.

جدول (۴): تلفیق ویژگی‌های انتشاری و ویژگی‌های به‌دست‌آمده از قوانین آپریوری

Ntvdm	Setup Api	TSUser Enabled
SHCore	Tiptsf	Msvbvm60
Monitors	SendMessage	True Type Font
Terminal	Image file extention	ws2_32
Create Remote Thread	netapi32	comdlg32
cfgmgr32	Profapi	Full Screen
Disable8And16BitMitigation		Cfgmgr 32

۴-۴- ایجاد بانک اطلاعاتی

در این مقاله یک دسته از بدافزارها به نام کرم‌ها که شامل کرم ایمیل، کرم پیام، کرم اینترنت، کرم شبکه نظیر به نظیر به همراه تعدادی از فایل‌های سالم موردبررسی قرار می‌گیرد. از هرکدام از انواع کرم‌ها (اینترنت، ایمیل، پیام، نظیربه‌نظیر) چندین نمونه در جعبه شنی اجرا می‌شود. پس از اجرای هر فایل نصبی در جعبه شنی یک فایل متنی شامل لاگ‌های مربوط به نصب و اجرای برنامه حاصل می‌شود. ویژگی‌های استخراج‌شده در مرحله قبل به‌عنوان سطر اول جدول در نظر گرفته می‌شود. به‌منظور تکمیل سطرهای بعدی جدول، هر فایل متنی به‌عنوان یک سطر از جدول قرار می‌گیرد. سپس وجود هر یک از ویژگی‌ها در فایل متنی بررسی می‌شود. در صورت پیدا شدن ویژگی در فایل متنی،ویژگی مقدار یک و در غیر این صورت مقدار صفر می‌گیرد. از طریق ادامه جست‌وجو ویژگی‌ها برای فایل‌های سالم و مخرب

پایگاه داده مربوطه ایجاد می‌شود. در این مرحله سه بانک اطلاعاتی برای هر سه دسته ویژگی ایجاد می‌شود.

۴-۵- ایجاد بانک اطلاعاتی:

در این مقاله یک دسته از بدافزارها به نام کرم‌ها که شامل کرم ایمیل، کرم پیام، کرم اینترنت، کرم شبکه نظیر به نظیر به همراه تعدادی از فایل‌های سالم موردبررسی قرار می‌گیرد. از هرکدام از انواع کرم‌ها (اینترنت، ایمیل، پیام، نظیربه‌نظیر) چندین نمونه در جعبه شنی اجرا می‌شود. پس از اجرای هر فایل نصبی در جعبه شنی یک فایل متنی شامل لاگ‌های مربوط به نصب و اجرای برنامه حاصل می‌شود. ویژگی‌های استخراج‌شده در مرحله قبل به‌عنوان سطر اول جدول در نظر گرفته می‌شود. به‌منظور تکمیل سطرهای بعدی جدول، هر فایل متنی به‌عنوان یک سطر از جدول قرار می‌گیرد. سپس وجود هر یک از ویژگی‌ها در فایل متنی بررسی می‌شود. در صورت پیدا شدن ویژگی در فایل متنی، ویژگی مقدار یک و در غیر این‌صورت مقدار صفر می‌گیرد. از طریق ادامه جست‌وجو ویژگی‌ها برای فایل‌های سالم و مخرب پایگاه داده مربوطه ایجاد می‌شود. در این مرحله سه بانک اطلاعاتی برای هر سه دسته ویژگی ایجاد می‌شود.

۴-۶- استفاده از الگوریتم‌های داده‌کاوی در طبقه‌بندی کرم‌ها:

در این مرحله، از الگوریتم‌های مناسب در ابزار وکا به‌منظور طبقه‌بندی داده‌هایی که در مرحله قبل به‌دست‌آمده است، استفاده می‌کند. از الگوریتم‌های مدنظر در ابزار وکا به‌منظور طبقه‌بندی داده‌ها استفاده می‌شود. در پایان این مرحله مدل یادگیری ایجاد می‌شود.

۴-۷- آزمون مدل یادگیری شده و تشخیص فایل ناشناخته

در مرحله آخر مدل یادگیری شده به‌منظور شناسایی و طبقه‌بندی فایل ناشناخته مورد آزمون قرار می‌گیرد. بدین منظور الگوریتم پیشنهادی فایل متنی حاصل از اجرای فایل ناشناخته در جعبه شنی را از ورودی می‌گیرد سپس ویژگی‌های موجود در جدول اولیه را با همان ترتیب در فایل متنی جست‌وجو می‌کند و فایل Arff مربوط به آن را ایجاد می‌کند. در پایان، این فایل Arff به مدل یادگیری به‌عنوان ورودی داده می‌شود سپس مدل یادگیری نوع فایل را مشخص می‌کند.

۵- نتایج و ارزیابی

این بخش به ارزیابی روش پیشنهادی اختصاص داده‌شده است. در این بخش، بعد از مطرح کردن پارامترهای ارزیابی و آزمون‌های انجام‌شده بر روی ویژگی‌های مختلف ارائه‌شده است.

۵-۱- معیارهای ارزیابی:

به‌منظور تحلیل و ارزیابی مدل طبقه‌بندی از پارامترهای نرخ مثبت کاذب[1]، نرخ مثبت درست[2]، ROC، دقت طبقه‌بندی مجذور میانگین مربعات خطا[3] و میانگین مطلق خطا[4] استفاده می‌شود.

- **نرخ مثبت کاذب:** یکی از اساسی‌ترین معیارهای یک سیستم تشخیص نفوذ ایده‌آل، به دست آوردن نرخ مثبت کاذب پایین است. منظور از نرخ مثبت کاذب این است که یک فایل سالم به‌عنوان یک فایل مخرب تلقی شود. نرخ مثبت کاذب به‌صورت FP نمایش داده می‌شود.
- **نرخ مثبت درست:** برابر است با درصدی از بدافزارها که به‌صورت مخرب شناخته شده‌اند.
- **مجذور میانگین مربعات خطا و میانگین مطلق خطا:** نشان‌دهنده میزان خطای مدل می‌باشند؛ که بهترین مقدار آن‌ها برابر صفر است و از طریق روابط (۱) و (۲) محاسبه می‌شوند.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{k=1}^{K}(x_k - y_k)^2}{Kk}} \qquad (۱)$$

$$\mathrm{MAE} = \frac{\sum_{k=1}^{k}|x_k| - |Y_k|}{k} \qquad (۲)$$

در روابط فوق x_k مقادیر مشاهداتی، y_k مقادیر برآورده شده و k تعداد داده‌ها است.

۶- نتایج آزمایش‌ها

در این مرحله، کارایی روش پیشنهادی مورد آزمون قرار داده می‌شود. به‌منظور تحلیل رفتار کرم‌ها و ایجاد پایگاه داده، از چهار خانواده کرم واقعی و یک مجموعه ۷۹ تایی از برنامه‌های سالم استفاده شده است. جدول (۵) مشخصات این چهار خانواده را توصیف می‌کند.

معیار انتخاب این چهار خانواده بر اساس مقالات ارائه‌شده در این زمینه است. منبع اصلی ما برای جمع‌آوری کرم‌های مربوطه هر خانواده سایت vx-archiv.at بوده است. در این راستا به‌منظور ارزیابی از مجموعه متنوع از برنامه‌های سالم که شامل ۱۳۲ نمونه است، استفاده شده است. بیش از ۹۰٪ از این برنامه‌های سالم به برنامه‌های شبکه و دسترسی از راه دور اختصاص داده شده است.

به‌منظور پیش‌پردازش داده‌ها بر روی هر سه دسته پایگاه داده حاصل‌شده، الگوریتم بازنمونه‌گیری[5] اعمال می‌شود. سپس برای ایجاد مدل یادگیری درصد تقسیم داده‌ها بر روی ۸۵٪ قرار

1- False Positive
2- True Positive
3- Root mean square error
4- Mean absolute error
5- Resample

می‌گیرد. این بدین معنی است که ۸۵٪ داده‌ها متعلق به یادگیری و ۱۵٪ باقیمانده مربوط به آزمودن روش پیشنهادی است.

الگوریتم پیشنهادی را بر روی پایگاه دادگان ساخته‌شده، مورد آزمایش قرار داده‌ایم. سپس ۴ پارامتر اندازه‌گیری شامل نرخ مثبت کاذب، نرخ مثبت درست و دقت طبقه‌بندی برای ارزیابی این روش استفاده کردیم. دقت به‌دست‌آمده از الگوریتم پیشنهادی با استفاده از چهار الگوریتم Random Forest، J48، Naïve Bayes، IB1 با در نظر گرفتن برنامه‌های سالم را در جدول (۷) ارائه‌شده است. دقت الگوریتم‌های Random Forest، J48، Naïve Bayes، IB1 در تشخیص و آزمون کرم‌ها با در نظر گرفتن فایل‌های سالم و بر اساس ویژگی‌های انتشاری به ترتیب ۱۰۰، ۹۳/۳۳، ۹۳/۳۳، ۱۰۰ است. از بین این چهار الگوریتم‌های Random Forest و IB1 بیش‌ترین دقت و الگوریتم‌های Naïve Bayes و J48 کم‌ترین دقت را دارد.

دقت به‌دست‌آمده از الگوریتم پیشنهادی با استفاده از چهار الگوریتم Random Forest، J48، Naïve Bayes، IB1 بدون در نظر گرفتن برنامه‌های سالم در جدول (۷) ارائه شده است. دقت الگوریتم‌های Random Forest، J48، Naïve Bayes، IB1 در تشخیص و آزمون کرم‌ها بدون در نظر گرفتن برنامه‌های سالم و بر اساس ویژگی‌های انتشاری به ترتیب ۶۶/۶۶، ۶۶/۶۶، ۷۳/۳۳، ۸۸/۶۶ است. از بین این چهار الگوریتم، الگوریتم IB1 بیش‌ترین دقت و الگوریتم‌های Naïve Bayes و Random Forest کم‌ترین دقت را دارد. سپس به‌منظور نشان دادن دقت ویژگی‌های انتشاری، الگوریتم پیشنهادی بر روی پایگاه داده ایجادشده از الگوریتم آپریوری و ترکیب دودسته ویژگی استخراج‌شده اعمال می‌شود.

دقت به‌دست‌آمده از الگوریتم پیشنهادی با استفاده از چهار الگوریتم Random Forest، J48، Naïve Bayes، IB1 با در نظر گرفتن برنامه‌های سالم و بر اساس ویژگی‌های آپریوری در جدول (۱۰) ارائه شده است. دقت الگوریتم‌های Random Forest، J48، Naïve Bayes، IB1 در تشخیص و آزمون کرم‌ها با در نظر گرفتن برنامه‌های سالم و بر اساس ویژگی‌های آپریوری به ترتیب ۹۶/۶۶، ۹۰، ۹۰، ۹۳/۳۳ است. از بین این چهار الگوریتم، الگوریتم RandomForest بیش‌ترین دقت و الگوریتم‌های Naïve Bayes و J48 کم‌ترین دقت را دارد.

دقت به‌دست‌آمده از الگوریتم پیشنهادی با استفاده از چهار الگوریتم Random Forest، J48، Naïve Bayes، IB1 بدون در نظر گرفتن برنامه‌های سالم و بر اساس ویژگی‌های آپریوری در جدول (۹) ارائه‌شده است. دقت الگوریتم‌های Random Forest، J48، Naïve Bayes، IB1 در تشخیص و آزمون کرم‌ها بدون در نظر گرفتن برنامه‌های سالم و بر اساس ویژگی‌های آپریوری به ترتیب ۷۳/۳۳، ۷۳/۳۳، ۶۰، ۶۶/۶۶ است. از بین این چهار الگوریتم، الگوریتم‌های Random Forest و J48 بیش‌ترین دقت و الگوریتم Naïve Bayes کم‌ترین دقت را دارد.

جدول (۵): توصیف مربوط به هر یک از چهار خانواده کرم استفاده‌شده

نام خانواده	توصیف	تعداد
ایمیل	انتشار از طریق ایمیل	۳۲
پیام	انتشار از طریق پیام	۳۱
نظیر به نظیر	به اشتراک‌گذاری فایل‌ها در شبکه نظیر به نظیر	۷
اینترنت	انتشار از طریق شبکه اینترنت	۳۱

جدول (۶): دقت طبقه‌بندی داده‌ها و معیارها مطرح‌شده روش پیشنهادی بر اساس بانک اطلاعاتی به‌دست‌آمده از ویژگی‌های انتشاری استخراج‌شده با در نظر گرفتن برنامه‌های سالم

نام الگوریتم	میانگین مطلق خطا	مجذور میانگین مربعات خطا	دقت طبقه‌بندی	نرخ مثبت درست	نرخ مثبت کاذب
Random Forest	۰	۰	۱۰۰	۱	۰
J48	۰/۰۳۵۹	۰/۰۱۳۵۶	۹۳/۳۳۳۳	۰/۹۳۳	۰/۰۳۷
Naïve Bayes	۰/۲۹۱	۰.۱۶۱۲	۹۳/۳۳۳۳	۰/۹۳۳	۰/۰۳۷
IB1	۰	۰	۱۰۰	۱	۰

جدول (۷): دقت طبقه‌بندی داده‌ها و معیارها مطرح‌شده روش پیشنهادی بر اساس بانک اطلاعاتی به‌دست‌آمده از ویژگی‌های انتشاری استخراج‌شده بدون در نظر گرفتن برنامه‌های سالم

نام الگوریتم	میانگین مطلق خطا	مجذور میانگین مربعات خطا	دقت طبقه‌بندی	نرخ مثبت درست
Random Forest	۰/۱۵۱۳	۰/۳۱۷۹	۶۶/۶۶۶۷	۰/۶۶۷
J48	۰/۱۷۴۹	۰/۳۴۳۹	۶۶/۶۶۶۷	۰/۶۶۷
Naïve Bayes	۰/۱۶۰۵	۰/۳۰۸۵	۷۳/۳۳۳۳	۰/۷۳۳
IB1	۰/۰۶۶۷	۰/۲۵۸۲	۸۸/۶۶۷۰	۰/۸۶۷

جدول (۸): دقت طبقه‌بندی داده‌ها و معیارها مطرح‌شده روش پیشنهادی بر اساس بانک اطلاعاتی به‌دست‌آمده از ویژگی‌های استخراج‌شده از آپریوری با در نظر گرفتن فایل سالم

نام الگوریتم	میانگین مطلق خطا	مجذور میانگین مربعات خطا	دقت طبقه‌بندی	نرخ مثبت درست	نرخ مثبت کاذب
Random Forest	۰/۰۳۵	۰/۱۳۵۳	۹۶/۶۶	۰/۹۶۷	۰/۰۳۳
J48	۰/۰۶۷۵	۰/۱۹۳۱	۹۰	۰/۹	۰/۰۷۳
Naïve Bayes	۰/۰۵۰۸	۰/۱۸۰۵	۹۰	۰/۹	۰/۰۷۳
IB1	۰/۰۲۶۷	۰/۱۶۳۳	۹۳/۳۳	۰/۹۳۳	۰/۰۶۷

جدول (۹): دقت طبقه‌بندی داده‌ها و معیارها مطرح‌شده روش پیشنهادی بر اساس بانک اطلاعاتی به‌دست‌آمده از ویژگی‌های استخراج‌شده از آپریوری بدون در نظر گرفتن فایل سالم

نام الگوریتم	میانگین مطلق خطا	مجذور میانگین مربعات خطا	دقت طبقه‌بندی	نرخ مثبت درست	نرخ مثبت کاذب
Random Forest	۰/۱۷۰۴	۰/۳۲۰۷	۷۳/۳۳	۰/۷۳۳	۰/۱۱۵
J48	۰/۱۸۱۱	۰/۳۱۸۷	۷۳/۳۳	۰/۷۳۳	۰/۱۱۵
Naïve Bayes	۰/۱۹۸۴	۰/۳۳۳۱	۶۰	۰/۶	۰/۱۱۵
IB1	۰/۱۶۶۷	۰/۴۰۸۲	۶۶/۶۶	۰/۶۶۷	۰/۱۳۲

دقت طبقه‌بندی داده‌ها و معیارها مطرح‌شده روش پیشنهادی بر اساس بانک اطلاعاتی به‌دست‌آمده از تلفیق دودسته ویژگی انتشاری و آپریوری با در نظر گرفتن برنامه‌های سالم با استفاده از چهار الگوریتم Random Forest، J48 ،Naïve Bayes، IB1 در جدول (۱۰) ارائه‌شده است. دقت الگوریتم‌های Random Forest، J48 ،Naïve Bayes، IB1 در تشخیص و آزمون کرم‌ها، با در نظر گرفتن برنامه‌های سالم و بر اساس بانک اطلاعاتی به‌دست‌آمده از تلفیق دودسته ویژگی انتشاری و آپریوری به ترتیب ۱۰۰، ۹۳/۳۳، ۸۳/۳۳، ۹۳/۳۳ است. از بین این چهار الگوریتم، الگوریتم Random Forest بیش‌ترین دقت و الگوریتم Naïve Bayes کم‌ترین دقت را دارد.

دقت طبقه‌بندی داده‌ها و معیارها مطرح‌شده روش پیشنهادی بر اساس بانک اطلاعاتی به‌دست‌آمده از تلفیق دودسته ویژگی انتشاری و آپریوری با استفاده از چهار الگوریتم Random Forest، J48 ،Naïve Bayes، IB1 بدون در نظر گرفتن برنامه‌های سالم در جدول (۱۱) ارائه شده است. دقت الگوریتم‌های Random Forest، J48 ،Naïve Bayes، IB1 در تشخیص و آزمون کرم‌ها بدون در نظر گرفتن برنامه‌های سالم و بر اساس بانک اطلاعاتی به‌دست‌آمده از تلفیق دودسته ویژگی انتشاری و آپریوری به ترتیب ۶۶/۶۶، ۶۶/۶۶، ۷۳/۶۰ است. از بین این چهار الگوریتم، الگوریتم IB1 بیش‌ترین دقت و الگوریتم Naïve Bayes کم‌ترین دقت را دارد.

جدول (۱۰): دقت طبقه‌بندی داده‌ها و معیارها مطرح‌شده روش پیشنهادی بر اساس بانک اطلاعاتی به‌دست‌آمده از تلفیق دودسته ویژگی انتشاری و آپریوری با در نظر گرفتن فایل سالم

نام الگوریتم	میانگین مطلق خطا	مجذور میانگین مربعات خطا	دقت طبقه‌بندی	نرخ مثبت درست	نرخ مثبت کاذب
Random Forest	۰/۰۲۴	۰/۰۸۶۳	۱۰۰	۱	۰
J48	۰/۰۴۵۸	۰/۱۵۷۵	۹۳/۳۳	۰/۹۳۳	۰/۰۳۴
Naïve Bayes	۰/۰۶۰۶	۰/۲۲۵۷	۸۳/۳۳	۰/۸۳۳	۰/۰۴۸
IB1	۰/۰۲۶۷	۰/۱۶۳۳	۹۳/۳۳	۰/۹۳۳	۰/۰۰۹

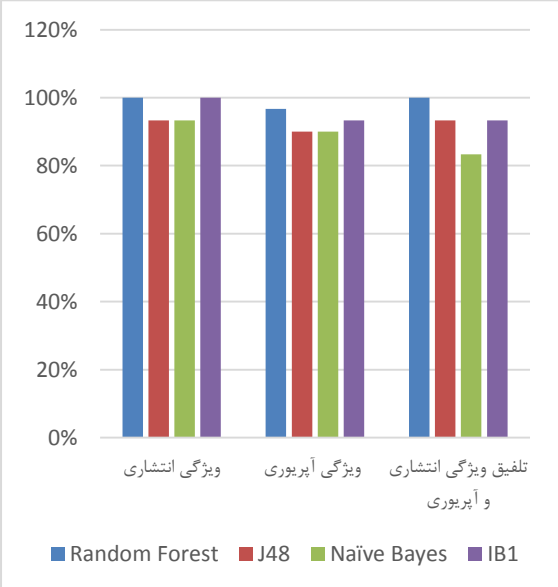
جدول (۱۱): دقت طبقه‌بندی داده‌ها و معیارها مطرح‌شده روش پیشنهادی بر اساس بانک اطلاعاتی به‌دست‌آمده از تلفیق دودسته ویژگی انتشاری و آپریوری بدون در نظر گرفتن فایل سالم

نام الگوریتم	میانگین مطلق خطا	مجذور میانگین مربعات خطا	دقت طبقه‌بندی	نرخ مثبت درست	نرخ مثبت کاذب
Random Forest	۰/۱۶۲۲	۰/۳۴۲	۶۶/۶۶	۰/۶۶۷	۰/۱۵۸
J48	۰/۱۷۴۹	۰/۳۴۳۹	۶۶/۶۶	۰/۶۶۷	۰/۱۵۸
Naïve Bayes	۰/۱۷۷	۰/۳۴۱۶	۶۰	۰/۶	۰/۱۲۳
IB1	۰/۱۳۳۳	۰/۳۶۵۱	۷۳/۳۳	۰/۷۳۳	۰/۱۰۸

آزمایش‌ها انجام‌شده نشان می‌دهد که مدل پیشنهادی با استفاده از ویژگی‌های انتشاری می‌تواند داده‌ها را با نرخ تشخیص بالایی طبقه‌بندی کند و فایل‌های ناشناخته را دسته‌بندی کند. نتایج این آزمایش‌ها در جدول‌های (۶)، (۸) و (۱۰) با در نظر گرفتن برنامه‌های سالم و در جدول‌های (۷)، (۹) و (۱۱) بدون در نظر گرفتن برنامه‌های سالم نشان داده شده است. بطوریکه دقت تشخیص و طبقه‌بندی داده‌ها با در نظر گرفتن برنامه‌های سالم با استفاده از الگوریتم‌های Random Forest، J48 Naïve Bayes، IB1، به ترتیب ۱۰۰، ۹۳/۳۳، ۹۳/۳۳، ۱۰۰ است. درحالی‌که در مرجع [۱۲] نرخ تشخیص داده‌ها بر اساس سه الگوریتم Naïve Bayes، IB1، j48 بر اساس دو مجموعه داده به ترتیب ۲۷/۵۰۹۹، ۶۹/۶۳۲۲، ۷۲/۲۶۶۴ و ۸۵/۵۳، ۹۶/۷۱، ۹۶/۰۷ درصد بوده است.

با توجه به‌دقت به‌دست‌آمده از این دودسته ویژگی می‌توان نتیجه گرفت که ویژگی‌هایی که بر اساس رفتار انتشار استخراج‌شده‌اند در تشخیص کرم‌ها به‌صورت مؤثرتری عمل می‌کنند. به‌گونه‌ای که دقت تشخیص و طبقه‌بندی داده‌ها با استفاده از الگوریتم‌های Random Forest، J48 Naïve Bayes، IB1، بر اساس ویژگی‌های انتشاری به ترتیب، ۱۰۰، ۹۳/۳۳، ۹۳/۳۳، ۱۰۰ درحالی‌که دقت تشخیص این چهار الگوریتم بر اساس ویژگی‌هایی که با استفاده از قانون آپریوری استخراج‌شده است به ترتیب ۹۶/۶۶، ۹۰، ۹۰ و ۹۳/۳۳ درصد است. این نشان می‌دهد که روش پیشنهادی بر اساس ویژگی‌های انتشاری می‌تواند برنامه‌های ناشناخته را با دقت بالاتری تشخیص دهد. از جدول‌های (۱۱-۱۰) می‌توان نتیجه گرفت که اگر ویژگی‌هایی که دستی استخراج شده‌اند را به ویژگی‌های به‌دست‌آمده از الگوریتم آپریوری اضافه کرد، دقت تشخیص کرم‌ها افزایش می‌یابد. الگوریتم جنگل تصادفی درروش پیشنهادی بر اساس ویژگی‌های انتشاری با دقت بالاتری نسبت به سه الگوریتم دیگر عمل می‌کند.

در شکل (۱) دقت طبقه‌بندی داده‌ها بر اساس سه دسته ویژگی انتشاری، آپریوری و تلفیق ویژگی انتشاری و آپریوری، نشان داده شده است. با توجه به این شکل می‌توان نتیجه گرفت که الگوریتم پیشنهادی بر اساس ویژگی‌های انتشاری توانسته است با دقت بالاتری در طبقه‌بندی داده‌ها عمل کند. به‌گونه‌ای که دقت به‌دست‌آمده بر اساس ویژگی‌های انتشاری ۱۰۰٪ است درحالی‌که دقت طبقه‌بندی داده‌ها بر اساس ویژگی‌های آپریوری ۹۳/۳۳۳۳٪ است. دقت به‌دست‌آمده از ویژگی‌های تلفیقی به‌دست‌آمده از ویژگی‌های انتشاری و آپریوری ۱۰۰٪ است؛ بنابراین، می‌توان نتیجه گرفت که ویژگی‌های انتشاری موجود در ویژگی‌های تلفیقی دقت طبقه‌بندی داده‌ها بر اساس ویژگی‌های تلفیقی افزایش داده است.



شکل (۱): دقت طبقه‌بندی داده‌ها با استفاده از هر سه دسته ویژگی استخراج‌شده

۷- نتیجه‌گیری

در این مقاله، یک روش جدید به‌منظور تشخیص و طبقه‌بندی کرم‌ها، مبتنی‌بر رفتار انتشار آن‌ها پیشنهاد شده است. ما توابع فراخوانی‌ای که کرم‌ها برای انتشار از آن‌ها استفاده می‌کنند را دسته‌بندی کردیم و آن‌ها را به‌عنوان ویژگی‌های شناسایی خانواده‌های کرم‌ها در نظر گرفتیم برای آزمون دقت این ویژگی‌های انتشاری از روش‌های یادگیری ماشین استفاده کردیم. به‌منظور اثبات صحت ویژگی‌های استخراج‌شده از الگوریتم آپریوری به‌منظور استخراج ویژگی از فایل‌های متنی استفاده شده است. ویژگی‌های انتشاری توانسته‌اند با دقت بیشتری نسبت به ویژگی‌های آپریوری عمل کنند. همچنین الگوریتم جنگل تصادفی و IB1 توانسته‌اند با دقت ۱۰۰ درصد نسبت به الگوریتم‌های Naïve Bayes ،J48 در تشخیص و طبقه‌بندی داده‌ها درروش پیشنهادی بر اساس ویژگی‌های انتشاری بهتر عمل کنند.

۸- منابع

1. P. M. Comparetti, G. Salvaneschi, E. Kirda, C. Kolbitsch, C. Kruegel, and S. Zanero, “Identifying Dormant Functionality in Malware Programs,” in IEEE Symposium on Security and Privacy, 2010 .
2. Y. Ye, T. Li, K. Huang, Q. Jiang, and Y. Chen, “Hierarchical associative classifier (HAC) for malware,” Journal of Intelligent Information Systems, vol. 35, no. 1, pp. 1-20, 2010.
3. N. R. Veeramani R, “Windows API based Malware Detection and Framework Analysis,” International conference on networks and cyber security, 2012 .
4. C. Raymond, S. Mancoridis, and M. Kam, “System Call-based Detection of Malicious Processes,” in IEEE International Conference on Software Quality, Reliability and Security, 2015 .
5. S. Naval, V. Laxmi, N. Gupta, M. Singh Gaur, and M. Rajarajan, “Exploring Worm Behaviors using DTW,” in Proceedings of the 7th International Conference on Security of Information and Networks, 2014 .
6. T. Barhoom and H. Qeshta, “Worm Detection by Combination of Classification with neural networks,” in international Arab Journal of e-Technology, 2013 .
7. D. Kharche and A. Thakare, “Internet Worm Classification and Detection using Data Mining techniques,” in IOSR Journal of Computer Engineering (IOSR- JCE), 2015 .
8. S. C. Jeeva and E. Blessing Rajsingh, “Intelligent phishing url detection using association rule mining,” in Human-centric Computing and Information Sciences, 2016.
9. Y. Wang, B. Watson, J. Zheng, and S. Mukkamala, “ARP-Miner: Mining Risk Patterns of Android Malware,” International Workshop on Multi-disciplinary Trends in Artificial Intelligence, pp. 363-375, 2015.
10. S. Palahan, D. Babić, S. Chaudhuri, and D. Kifer, “Extraction of Statistically Significant Malware Behaviors,” in Proceedings of the 29th Annual Computer Security Applications Conference, 2013 .
11. Y. Tang, J. Luo, B. Xiao, and G. Wei, “Concept, characteristics and defending mechanism of worms,” in IEICE transactions on Information and Systems, 2009 .
12. M. Norouzi, A. Souri, and M. Samad Zamini, “A Data Mining Classification Approach for Behavioral Malware Detection,” Journal of Computer Networks and Communications, 2016.

Analysis and Detection of Worm Propagation Behavior

M. H. Alaeiyan, SH. Sadeghnia, S. Parsa*

Abstract

The number of malwares, which is the most important communication networks security challenge, increases quickly. Malware cause financial and physical harm to individuals and organizations. Worms are a type of malware which spread through emails, p2p networks and Internet connections automatically. Therefore, identifying various propagation behaviors of worms helps us to classify them. To classify worms, both benign and malicious programs are executed within a sandbox to screen API calls. Thus, we analyze the sequence of extracted API calls to derive propagation behaviors. Propagation behaviors are defined as propagation features. Further, random forest algorithm classified worms' families with an accuracy of 100% while the features which are obtained by Aprior algorithm classified worm families with an accuracy of 96.66%..

Key Words: *Worm Detection, Propagation Behaviors, Sequence of API calls*

* Iran University of Science and Technology (Parsa@iust.ac.ir)- Writer-in-Charge