نشریه علمی "پدافند الکترونیکی و سایبری"
سال هفتم، شماره ۴، زمستان ۱۳۹۸، ص ۷۷-۶۷

تشخیص هوشمندکرم‌های نهان در شبکه‌های رایانه‌ای

مهیار ذبیحی[۱]، بهروز مینایی[۲*]، مهدی نصیری[۳]
۱- کارشناس ارشد دفاع سایبری، دانشگاه جامع امام حسین(ع)، ۲- دانشیار، ۳- دکترای داده‌کاوی، دانشگاه علم و صنعت ایران

(دریافت: ۹۶/۱۱/۲۶، پذیرش: ۹۷/۰۷/۲۱)

چکیده

یکی از تهدیدات اساسی درفضای سایبر، بدافزارهای پیچیده‌ای می‌باشد که به قصد جاسوسی و تخریب سامانه‌ها در زیرساخت‌های حیاتی کشور گسترش یافته است. آنچه در این مقاله ارائه‌شده روشی هوشمند در کشف کرم‌های نهانی می‌باشد که می‌تواند چندریختی و رمزشده بوده و ماهیت آنها هنوز برای ابزارهای دفاعی ناشناخته باقی مانده است. برای این منظور با تأکید بر ویژگی‌های پویش کرم، مدل ارتباطات میزبان‌های آلوده و سرآیند بسته‌های ارسالی روی بستر شبکه، راه‌کاری مبتنی بر روش‌های داده‌کاوی در کشف گسترش‌های مخرب ارائه نمودیم. با خوشه‌بندی داده‌های پاک و استفاده از رده‌بندی داده‌های پاک و آلوده و به‌کارگیری نمونه‌های آزمایشگاهی توانستیم بهترین مدل را به کمک روش درخت تصمیم C5 با صحت ٪ ۹۴/۴۹، دقت ٪۹۲/۹۲ و با بازخوانی ٪۹۴/۷۰ در کشف بسته‌های آلوده از پاک ارائه نماییم. در نهایت نیز نشان دادیم که استفاده از خوشه‌بندی در الگوهای ترافیک میزبان‌های پاک نتایج بهتری را در شناخت ترافیک‌های آلوده به‌دست می‌آورد.

کلید واژه‌ها: کرم، تشخیص‌دهنده، پویش، داده‌کاوی.

۱. مقدمه

امروزه اینترنت به یکی از مولفه‌های اصلی زندگی بشر تبدیل شده و در دسترس بودن و امنیت آن از اهمیت روزافزونی برخوردار است. انواع مختلفی از بدافزارها اعم از کرم‌ها، تروجان‌ها و ویروس‌ها عواملی هستند که این مهم را مورد تهدید قرار می‌دهند. بدیهی است برای مبارزه با این بدافزارها تشخیص آنها اولین گام در معماری ساختار دفاعی می‌باشد آنچه در این مقاله به آن پرداخته شده روشی برای تشخیص کرم‌های خودانتشاردهنده[۱] می‌باشد، که در طول سالیان اخیر روزبه‌روز پیچیده‌تر و هوشمندتر شده‌اند. در دهه‌های پیشین اگر امضای کرم مشخص می‌شد به‌سادگی در شبکه کشف می‌گردید اما کرم‌های امروزی حتی رفتار خود را در طی انتشار در شبکه تغییر می‌دهند. برای این منظور در مرحله نخست به بررسی انواع کرم‌ها و روش‌های تشخیص و مقابله با آنها خواهیم پرداخت سپس کارایی تشخیص‌دهندگان[۲] موجود کرم را مورد ارزیابی قرار می‌دهیم در این راه حل قصد داریم تا به کمک داده‌کاوی[۳] و استخراج ویژگی‌های مناسب از بسته‌های[۴] ترافیک شبکه یک تشخیص‌دهنده هوشمند به‌منظور تشخیص کرم‌های نهان در شبکه‌های رایانه‌ای ارائه نماییم در این مدل ضمن استخراج ویژگی‌های مورد نیاز در عملیات اصلی، تلفیقی از عملیات خوشه‌بندی و رده‌بندی را انجام داده و نتایج به‌دست‌آمده را مورد ارزیابی قرار می‌دهیم.

در ابتدای این مقاله در بخش دوم ساختار کرم‌های رایانه‌ای و روش‌های پویش[۵] آنها را بررسی می‌نماییم سپس در بخش سوم ضمن مطالعه ویژگی‌های تشخیص‌دهنده‌های مناسب تحقیقات مرتبط صورت‌پذیرفته در این خصوص را ارزیابی می‌نماییم. در این بخش روش‌شناسی تحقیق شامل روش استفاده از سرآیندهای بسته‌های ترافیک رمزشده به همراه مدل کلی تحقیق در به‌کارگیری روش‌های خوشه‌بندی و دسته‌بندی داده‌کاوی در کشف کرم‌ها ارائه می‌گردد بخش چهارم به پیاده‌سازی و استخراج ویژگی‌های مورد نیاز در عملیات داده‌کاوی می‌پردازد. در این قسمت ضمن متوازن‌سازی داده‌های آلوده به روش‌های تصادفی و خوشه‌بندی داده‌های پاک نتایج را در مرحله مدل‌سازی مورد مقایسه قرار می‌دهیم و با به‌کارگیری روش‌های رده‌بندی در تشخیص میزبان‌های آلوده و پاک و استفاده از سه روش ماشین بردار پشتیبان، درخت تصمیم و شبکه‌های عصبی مصنوعی نتایج را یک‌بار در وضعیت عدم استفاده از خوشه‌بندی و به‌کارگیری

* رایانامه نویسنده مسئول: b_minaei@iust.ac.ir

[1] Self-Propagating
[2] Worm Detectors
[3] Data mining
[4] Packets
[5] Scan

متوازن‌سازی تصادفی و بار دیگر در حالت به‌کارگیری خوشه‌بندی به‌دست آورده و آنها را ارزیابی نمودیم.

پیاده‌سازی خود را به کمک نرم‌افزار کلمنتاین انجام داده و برای یادگیری از ۸۰ درصد ترافیک و برای تست از ۲۰ درصد آن استفاده نمودیم. در بخش پنجم نیز نتایج تحقیق را ارزیابی و جمع‌بندی خواهیم نمود و در آن با بررسی سه روش رده‌بندی به‌کارگرفته شده بهترین نتایج که متعلق به درخت تصمیم C5 در کشف بسته‌های آلوده از پاک می‌باشد را ارائه خواهیم نمود.

۲. کرم‌های رایانه‌ای و روش‌های پویش آنها

بدافزارها به انواع متعددی تقسیم می‌گردند که یکی از آنها کرم‌ها می‌باشند. در واقع کرم‌ها برنامه‌ای نرم‌افزاری هستند که قادرند تا با به‌دست‌آوردن دسترسی‌های لازم، خود را برروی رایانه‌های دیگر تکثیر کنند و به کمک آسیب‌پذیری‌های موجود در سیستم میزبان، برنامه خود را اجرا نمایند. این آسیب‌پذیری‌ها یا در یکی از نرم‌افزارهای سیستم میزبان وجود دارد یا بر اثر پیکربندی ناصحیح نرم‌افزارهای سیستم میزبان به‌وجود آمده است.

به‌طور کلی فعالیت کرم‌ها به دو بخش یافتن قربانی و مورد اصابت قرار دادن آن صورت می‌پذیرد. برای این منظور کرم می‌بایست ابتدا سیستم قربانی را پیدا نموده و پس از کپی کردن کد مخرب، آن را اجرا نماید. بدیهی است که یکی از مهم‌ترین مراحل انتشار برای آنها پویش شبکه به‌منظور یافتن قربانیان خود می‌باشد آنچه در موفقیت پویشگر کرم تأثیر به سزایی دارد روشی است که کرم در یافتن میزبان‌های آسیب‌پذیر خود به‌کار می‌بندد. به‌عنوان مثال فرض کنید پویشگری در یک شبکه کاملاً تصادفی عملیات پویش را انجام می‌دهد. بنابراین، برای یافتن میزبان هدف خود می‌بایست مجموعه حالات $N = 2^{32}$ را در بدترین حالت پویش کند. اکنون فرض کنید دو بخش اول پیش‌آدرس شبکه (مثلا ۱۹۲,۱۶۸) را بداند در این صورت مجموعه حالات به $N = 2^{16}$ حالت کاهش می‌یابد واگر پویشگر به هر طریقی بتواند فهرست میزبان‌های آسیب‌پذیر مدنظر خود را به‌دست آورد زمان این پویش به مرتبه خطی تقلیل می‌یابد.

۳. تشخیص‌دهنده مناسب

وقتی صحبت از سیستم تشخیص کرم می‌گردد در وهله اول مهم‌ترین مسئله برای ما تشخیص صحیح است. تشخیص صحیح بدین معناست که تشخیص‌دهنده می‌بایست علاوه بر این‌که تمامی فایل‌های مخرب را تشخیص می‌دهد به هیچ وجه نباید فایلی را که کرم نباشد به‌عنوان کرم معرفی کند همچنین زمان تشخیص نباید بیش از حد طولانی باشد و یا منابع زیادی از سیستم را در اشغال خود قرار دهد علاوه بر این بهتر است تشخیص‌دهنده به‌سادگی بر روی میزبان‌های پایانی شبکه انتشار یابد و کمترین سربار در نصب آن ایجاد شود بنابراین، بهتر است تشخیص‌دهنده در دروازه شبکه مستقر گردد و عملیات تشخیص را در این نقطه انجام دهد.

۳-۱. روش‌های تشخیص کرم

به‌طور کلی روش‌های تشخیص کرم از یک دیدگاه در دو دسته کلی تشخیص بر روی میزبان و تشخیص از روی بسته‌های جابجا شونده در بستر شبکه صورت می‌پذیرد و از دیدگاه دیگری نیز به دو دسته مبتنی بر امضا یا شکل کد بدافزار و تشخیص بر مبنای رفتار بدافزار تقسیم می‌گردند. تمرکز بر شکل بدافزار ممکن است با توجه به رمزبودن کد آن و یا دگردیسی آن در جابجایی بین دو میزبان عملیات تشخیص را با مشکل مواجه نماید همچنین قابلیت مهندسی معکوس و توانایی تحلیل و ارائه مدل‌های داده‌کاوی بر روی قطعات کد می‌تواند کمک شایانی به تشخیص از روی کد نماید اما از آنجا که موضوع این تحقیق بر روی کرم-های نهانی است که خود را در بستر شبکه رمز می‌نمایند پس این روش نمی‌تواند کارایی چندانی برای ما داشته باشد. از منظر دیگر تحلیل بر مبنای رفتار نیز مشکلات خاص خود را دارد. مهم‌ترین مسئله پنهان نمودن رفتار مخرب تا مقصد نهایی و شناسایی محیط هدف می‌باشد که کشف بر مبنای رفتار را با مشکل جدی مواجه می‌نماید. حتی در برخی موارد کرم‌ها قادرند تا محیط‌های آزمایشگاهی، جعبه شن[1] و سیستم‌عامل‌های مجازی را تشخیص داده و هیچ رفتار مخربی را از خود بروز ندهند. بنابراین، می‌بایست با ارائه یک روش رفتاری غیرمتمرکز بر رفتار نهایی کرم عملیات تشخیص را صورت بخشید.

کرم‌ها برای رسیدن به اهداف خود مجبورند خود را منتشر سازند همچنین می‌بایست از آسیب‌پذیری‌های موجود در سیستم‌عامل یا نرم‌افزارهای هدف استفاده نمایند بنابراین، تمرکز بر این دو نکته می‌تواند کلید اصلی راه‌حل کشف کرم باشد.

۳-۲. انتخاب تشخیص‌دهنده مطلوب

با توجه به آنچه در بخش قبلی گفته شد در این بخش می‌خواهیم تا قابلیت‌های تشخیص‌دهنده مطلوب خود را انتخاب کنیم. در تشخیص‌دهنده‌های میزبان محور یکی از اساسی‌ترین مشکلات مسئله نصب این تشخیص‌دهنده‌ها می‌باشد که اولاً ممکن است نتوانیم تمامی میزبان‌ها را مجبور به نصب کنیم و ثانیاً ممکن است به دلایلی صاحبان میزبان و یا بدافزاری خاص، نرم‌افزار

[1] Sandbox

نصب‌شده را غیرفعال نمایند. بنابراین، با وجود آن‌که در نقاط نهایی شبکه، دسترسی به حافظه اجرایی بیشتر است اما همواره این مشکل وجود دارد که میزبان مذکور آلوده بوده و نشر آلودگی صورت پذیرفته است بنابراین، ما روش‌های میزبان محور را از حوزه مطالعه خود کنار گذاشته و در ادامه تحقیق به روش‌های شبکه محور خواهیم پرداخت.

در روش‌های شبکه‌محور بخشی از این الگوریتم‌ها به دنبال الگوی توالی بایت‌های تکراری ترافیک مشکوک می‌گردند که مشکل آنها در تشخیص اجزای رمزشده کرم‌ها می‌باشد بنابراین، با توجه به این‌که تقریباً اکثر کرم‌های جدید چندریختی هستند و ترافیک خود را رمز می‌کنند این دسته را نیز از مطالعه خود خارج می‌نماییم.

دسته بعدی در روش‌های شبکه محور تشخیص‌دهنده‌هایی هستند که بر مبنای محتوای بسته‌های مرتبط با بخش‌های ثابت در کدکرم (که کرم مجبور است تا آنها را دست‌کاری نکند) عملیات تحلیل خود را صورت می‌بخشند. اینجا نیز مشکلی اساسی وجود دارد و آن این‌که پیدا کردن بخش ثابت و غیررمزشده در تمامی کدهای مخرب کاری دشوار است و اگر برای رفع این مشکل بخش ثابت را به شکل عمومی و وابسته به پروتکل ساخت تمامی کرم‌ها درنظر بگیریم با توجه به اشتراک زیادی که با برنامه‌های مجاز دارد مثبت کاذب را در شبکه بالا می‌برد همچنین ترافیک‌های مجاز رمزشده‌ای نیز وجود دارد که به اشتباه ممکن است کرم تشخیص داده شود. بنابراین، باید به دنبال روش شبکه محوری بگردیم که به‌صورت پویا و بر مبنای شاخص‌های رفتاری بدون تکیه بر محتوای بسته و با توجه به ترافیک میزبان‌های پایانی بتواند کرم‌های چندریختی را که در بستر رمز شده انتشار می‌یابند کشف نماید.

۳-۳. تحقیقات مرتبط

یکی از مهم‌ترین شیوه‌هایی که در سال‌های اخیر در جهت توسعه روش‌های مقابله با کرم‌های رایانه‌ای در حال گسترش می‌باشد به‌کارگیری مفاهیم داده‌کاوی در کشف کرم‌های روز صفرم می‌باشد در این راستا کارهای متعددی صورت پذیرفته که غالباً با تکیه بر ویژگی‌های اساسی ترافیک شبکه، خود را تکامل بخشیده است. در این رابطه بارتس و سافکا در سال ۲۰۱۵ طی پژوهشی از خصیصه‌های بسته‌های ارسالی بر شبکه استفاده نموده و با تمرکز بر محتوای آنها به کمک ترافیک غیررمزشده در بستر شبکه مخرب بودن منبع ارسال ترافیک را شناسایی نمودند [۱].

این تحقیق در سال ۲۰۱۷ توسط پراس و همکارانش تکامل بیشتری یافت. آنها با تمرکز بر ترافیک رمزشده روی بستر VPN و HTTPS ویژگی‌های اساسی مدنظر خود را استخراج نمودند. آنها به کمک داده‌کاوی و به‌کارگیری شبکه‌های عصبی بر روی نام دامنه‌های بسته‌ها و ویژگی‌های اندازه بسته‌های ورودی و خروجی و زمان ارسال بسته‌ها توانستند با دقت ۷۰ درصد در محیط واقعی بدافزارهای روی ترافیک رمزشده را تشخیص دهند [۲].

با بررسی دقیق‌تر این روش در می‌یابیم که مشکل اصلی روش آنها تکیه بر روی پروتکل HTTPS، نام دامنه و استفاده از N-gram در فراخوانی آدرس‌های دیگر میزبان‌ها می‌باشد زیرا دامنه کشف به بدافزارهایی محدود می‌شود که جهت انتشار خود از این پروتکل بهره می‌جویند همچنین ممکن است نام دامنه‌ای، در انتشار مطرح نبوده و بر پایه پویش فضای آدرس‌دهی پیرامونی، میزبان‌های قربانی کشف شود (مثلاً از اتصالات موجود در net stat استخراج شود). لازم به ذکر است که صرف تکیه بر اندازه بسته‌ها و زمان ارسال آنها، الگوی انتشار سریع را در نشر کرم تداعی می‌کند بنابراین، احتمالاً این روش نمی‌تواند در برابر انتشارهای با سرعت کمتر کارایی داشته باشد. مسئله دیگر نیز فرض ضمنی مسئله در چشم‌پوشی از عملیات یافتن قربانی در این نوع انتخاب ویژگی می‌باشد که می‌تواند کمک شایانی در تشخیص رفتار کرم نماید. در آخر نیز خاطر نشان می‌شویم که ترافیک مورد آزمایش از طریق جمع‌آوری بدافزارهای شناخته‌شده‌ای مورد ارزیابی قرار می‌گیرد که الزامی در استفاده از پروتکل HTTPS در انتشار خود رعایت نمی‌نماید بنابراین، نمی‌توان در این باره نظری قطعی بر روی بدافزارهای روز صفرم ارائه نمود.

۴-۳. روش‌شناسی تشخیص‌دهنده پیشنهادی

برای دستیابی به اسلوب مطلوب به‌منظور پوشش نقاط قوت و ضعف مطرح‌شده با استفاده از یادگیری ماشین و با الهام بخشی از ترکیبی از روش‌های تشخیص در شبکه اسلوب روش پیشنهادی را ایجاد می‌نماییم.

همان‌گونه که در قسمت قبل ذکر گردید داده‌کاوی در روش‌های تشخیص بسته‌ها بر روی شبکه با وجود ترافیک رمزشده می‌تواند نتایج مؤثری را در کشف کرم ایجاد نماید. ما نیز با نوآوری ارائه‌شده در ایجاد سه تفاوت نسبت به روش پیشنهادی پراس، روش تشخیص خود را ارائه می‌نماییم. اول اینکه روش ما منحصر به پروتکل یا روش رمزنگاری خاص نظیر HTTPS یا VPN نبوده و قادر است تا هرگونه ترافیک رمزشده‌ای بر بستر

TCP را ارزیابی نماید. ما در روش پیشنهادی به متن بسته کاری نداریم و صرفاً با اتکا به اطلاعات مشتق‌شده از فیلدهای سرآیند بسته‌ها نظیر آدرس‌های مبدأ و مقصد، طول عمر بسته و اندازه بسته عملیات خود را انجام می‌دهیم.

دوم اینکه بسته‌های مورد مطالعه ما می‌تواند صرفاً دارای یک آدرس مقصد باشد و هیچ نیازی به نام دامنه مقصد و یا اطلاعات سرآیند بسته‌های HTTP نمی‌باشد.

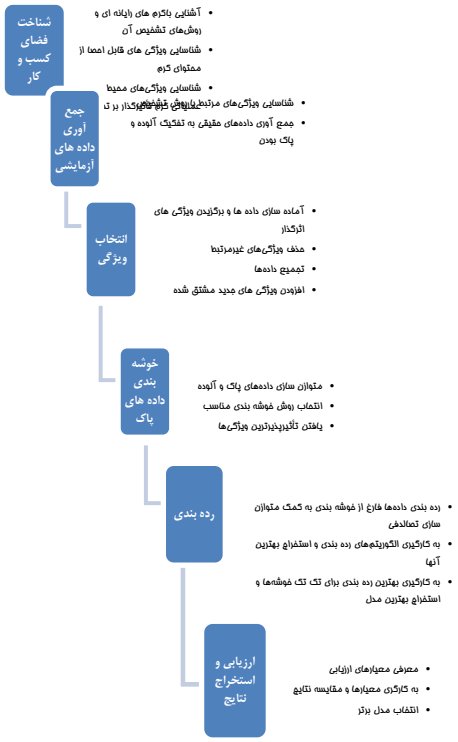
سوم هم روش داده‌کاوی ما می‌باشد که ترکیبی از عملیات خوشه‌بندی به همراه دسته‌بندی با روش‌های درخت‌های تصمیم، شبکه‌های عصبی و رگرسیون خطی می‌باشد که می‌تواند بهبود مؤثری در کشف این‌گونه بدافزارها ایجاد نماید.

برای انتخاب بردارهای ویژگی لازم است تا با الهام‌گیری از روش‌های تشخیص شبکه‌محور عملیات استخراج صورت پذیرد. با توجه به این‌که عمده تمرکز بر دو روش تشخیص عملیات پویش و رابطه علت و معلولی[1](پویش سلسله‌مراتبی) بین میزبان‌های آلوده قرارگرفته لازم است تا در انتخاب بردارهای ویژگی مقادیر مرتبط با وضعیت اتصالات قرار گیرد. برای این منظور می‌بایست این ویژگی از ویژگی‌های مرتبط با بسته‌های TCP استخراج شده و برای هر بسته با توجه به وضعیت کلی اتصال آن در طی یک گفتگو بین دو میزبان تصویر گردد. مسئله دیگر تلاش‌های ناموفق در یافتن هدف به‌دلیل ماهیت پویش کرم (در مواردی که فهرست از پیش تعیین شده‌ای در اختیار کد مخرب نیست) در تشخیص آن می‌باشد. برای این منظور از فراوانی فراخوانی‌های صورت پذیرفته توسط میزبان، طی اتصالاتش به دیگر میزبان‌ها استفاده شده و این ویژگی نیز در مرحله دوم از اتصالات موفق یا ناموفق به‌دست آمده تصویر شده و به مجموعه ویژگی‌ها اضافه می‌گردد. بحث بعدی الگوی علت و معلولی است که برای این منظور مجموعه ویژگی‌ها وارد سطح‌بندی شده و در سه سطح این الگو تسری می‌یابد. همچنین الهام‌گیری از الگوی آدرس مقصد در شکل‌گیری تعدد مراجعات با ایجاد یک شاخص جمعی روی درگاه و آدرس مقصد برای هر سطح مجموعه مراجعات محاسبه می‌گردد. البته لازم به ذکر است که استفاده از الگوهای حجمی نیز می‌تواند در تشخیص کرم‌هایی که دگردیسی ندارند کارا باشد همچنین به‌دست آوردن الگوی زمانی در شروع و پایان اتصالات بین میزبان‌ها نیز در تشخیص رفتار توزیع کرم و فاصله گرفتن آن از رفتار توزیع بسته‌های عادی بسیار مؤثر است.

[1] Causation

۳-۵. مدل کلی تحقیق

در این پژوهش از دو مدل خوشه‌بندی و رده‌بندی استفاده شده است. از خوشه‌بندی در نمونه‌گیری و ایجاد مجموعه داده‌های متنوع‌تر و از روش‌های رده‌بندی به‌منظور ساخت مدل‌های پیش‌بینی کننده استفاده شده است. در زیر مدل کلی تحقیق بر طبق مدل فرآیندی کریسپ کشیده شده و پیاده‌سازی آن با نرم‌افزار کلمنتاین[2] صورت پذیرفته است.



شکل (۱): مراحل پیاده‌سازی مدل تحقیق

۴. مدل‌سازی

در این بخش مدل‌سازی تحقیق طبق رویه تشریح شده در قسمت‌های قبلی، ارائه می‌گردد. برای این منظور در ابتدای امر مجموعه ویژگی‌های مورد استفاده معرفی می‌شود. در مرحله بعد عملیات تجمیع و کاهش بر روی آنها صورت پذیرفته و شاخصه‌های کلیدی جهت انجام عملیات داده‌کاوی استخراج می‌گردند. به کمک روش‌های خوشه‌بندی عملیات متوازن‌سازی صورت پذیرفته و با استفاده از رده‌بندی بهترین روش در کشف بسته‌های آلوده ارائه می‌گردد در انتهای این بخش نیز بر اساس به‌کارگیری خوشه‌بندی بر روی بسته‌های پاک و رده‌بندی آنها مدل، اصلاح شده و نتایج به کمک شاخصه‌های ارزیابی تشریح می‌شوند.

[2] Clementine

۴-۱. درک داده‌ها

یافتن یک محیط آزمایشگاهی واقعی مسئله‌ای بسیار مهم و کلیدی در ارزیابی مدل محسوب می‌گردد و از آنجایی که ما برای ارزیابی مدل خود نیازمند ترافیک میزبان‌های پاک و آلوده به کرمی هستیم که سعی در گسترش آلودگی خود در شبکه دارند و ایجاد این محیط مسئله‌ای بسیار دشوار می‌باشد از ترافیک ارائه‌شده توسط آزمایشگاه ICSI که ترافیک آلوده و پاک را به‌طور تفکیک‌شده طبقه‌بندی نموده استفاده کردیم [۳]. این آزمایشگاه متعلق به دانشگاه برکلی کالیفرنیا می‌باشد و ترافیک مورد استفاده ما در حدود ۵۰۱۰ میزبان عضو دارد. فایل‌های Pcap این ترافیک در آدرس مرجع [۳] قرار گرفته و فهرست میزبان‌های آلوده و پاک موجود می‌باشد. میزبان‌های آلوده اغلب دارای نرم‌افزارهای پویش شبکه، بدافزار و کرم‌های متنوع بوده و ترافیک آن در بازه زمانی یک روز ذخیره و مورد ارزیابی قرار گرفته است. این ترافیک مشتمل بر ۳۷۷۱۰۹۰ بسته داده تولیدشده توسط میزبان‌ها بوده که پس از کاهش اولیه به تعداد ۲۸۶۵۵۲۳ تقلیل یافته است. از این میان تعداد ۴۳۹ میزبان آلوده و مابقی آن پاک ارزیابی شده است. داده‌های موجود در این ترافیک شامل ویژگی‌های اولیه زمان اتصال، پروتکل، آدرس مبدأ، درگاه مبدأ، آدرس مقصد، درگاه مقصد، طول بسته، مدت زمان شروع اتصال تا زمان ارسال بسته دریافتی، فاصله زمانی بین دو بسته متوالی در یک اتصال، محتوای سرآیند و محتوای بسته می‌باشد که به‌دلیل رمز بودن پروتکل کرم نهان از محتوای بسته صرف‌نظر خواهیم نمود.

در اولین مرحله کلیه فایل‌ها تجمیع شده و در قالب یک جدول قرار می‌گیرد. هریک از ستون‌های این جدول یکی از فیلدهای اطلاعات سرآیند بسته‌ها می‌باشد و عملیات کاهش اولیه با جدا نمودن ترافیک مرتبط با پروتکل TCP اجرا می‌گردد. در مرحله بعد خصیصه اساسی وضعیت اتصال از مقادیر مرتبط با Rst , Syn , Ack مشتق شده با نام ConnectionStatus در مجموعه داده‌ها اضافه می‌گردد[۱] در این خصیصه وضعیت موفقیت یا عدم موفقیت در برقراری ارتباط توسط رکورد مرتبط با ترافیک مذکور قرار گرفته و برای به‌دست آوردن آن از مجموعه قواعد پروتکل TCP در برقراری ارتباط کمک گرفته شده است. بدین ترتیب که اگر در توالی ارسال بسته‌ها عملیات دست‌تکانی کامل شده باشد اتصال، موفق و در غیر این‌صورت ناموفق ارزیابی می‌گردد بدیهی است طی مطالبی که پیشتر ارائه شد این خصیصه می‌تواند کمک شایانی به تشخیص کرم بودن نماید.

در مرحله سوم تعداد اتصالات بین دو میزبان در بازه ترافیکی به‌دست می‌آید این مشخصه نیز در مفاهیم حرکت کرم‌های نهان جایگاه ویژه‌ای دارد شاید بتوان با احتمال زیادی بیان نمود که در صورت پویش توسط کرم اتصالات موفق یا ناموفق بین دو میزبان در الگوی خاصی قرار می‌گیرد. برای دستیابی به نتایج بهتر این تعداد را در چهار دسته فراوانی تک اتصالی‌ها، ۲ تا ۵ اتصالی‌ها، ۶ تا ۱۰ اتصالی‌ها و بیش از ۱۰ اتصالی‌ها تقسیم نمودیم. (این طریق دسته‌بندی بر مبنای خوشه‌بندی همراه با اصلاح فراوانی انواع اتصالات به‌دست می‌آید). این تعداد اتصالات می‌تواند موفق یا ناموفق باشد.

در مرحله چهارم برای هر میزبان و درگاه آن یک رکورد در جدول نهایی ایجاد نموده و از جدول ابتدایی مقادیر مجموع پنج خصیصه را محاسبه و در این جدول قرار دادیم این خصیصه‌ها به شرح زیر می‌باشد:

۱. وضعیت اتصال (ConnectionStatus):

مقدار آن می‌تواند صفر (به معنای اتصال غیر موفق) و یا یک (به معنای اتصال موفق) باشد. لازم به ذکر است که در صورتی که برای یک میزبان روی یک درگاه خاص هم اتصال موفق و هم ناموفق وجود داشت دو سطر از جدول را به خود اختصاص می‌دهد.

۲. تعداد رکوردهای تک اتصالی میزبان (One):

این خصیصه بیانگر مجموع تعداد اتصالات تکی می‌باشد که از منبع مذکور روی درگاه تعیین‌شده در وضعیت موفق (در صورت یک بودن وضعیت اتصال) و یا ناموفق (در صورت صفر بودن آن) به یک مقصد خاص صورت پذیرفته است. به‌عنوان مثال فرض کنید تنها یک اتصال موفق بین منابع A و B و یک اتصال موفق بین A و C و تعداد ۴ اتصال موفق بین منابع A و D وجود دارد بنابراین، مقدار این خصیصه برای منبع A برابر ۲ خواهد گردید.

۳. تعداد رکوردهای یک تا ۵ اتصالی میزبان(Plus01):

در این حالت تعداد اتصالات موفق یا ناموفق بین مبدأ و مقصد که بین یک تا ۵ بوده شمارش شده و در این خصیصه قرار می‌گیرد. در مثال بالا مقدار این ویژگی برای منبع A مقدار ۱ خواهد بود زیرا فقط یک ارتباط ۴ اتصالی بین A و D موجود می‌باشد.

---

۱ کلیه پرس و جوی‌ها در صورت نیاز از طریق پست الکترونیکی نویسنده اصلی قابل دریافت خواهد بود

۴. تعداد رکوردهای ۵ تا ۱۰ اتصالی میزبان(Plus05):
در این حالت تعداد اتصالات موفق یا ناموفق بین مبدأ و مقصد که بین ۵ تا ۱۰ بوده شمارش شده و در این خصیصه قرار می‌گیرد.

۵. تعداد رکوردهای بیش از ۱۰ اتصالی میزبان(Plus10):
در این حالت تعداد اتصالات موفق یا ناموفق بین مبدأ و مقصد که بیش از ۱۰ اتصال بوده شمارش شده و در این خصیصه قرار می‌گیرد.

در مرحله پنجم به هر یک از سطرهای مرحله قبلی مجموع اندازه بسته‌های مبادله شده توسط میزبان روی درگاه موردنظر و وضعیت ارتباط موفق یا ناموفق به‌عنوان خصیصه جدید افزوده می‌گردد.

همین مجموع در مورد خصیصه فاصله زمان بسته، نسبت به زمان آغازین اتصال به ازای هر بسته برای تمامی میزبان‌ها ایجاد گردیده است. همچنین مجموع فاصله‌های زمانی بسته‌های متوالی برای هر میزبان محاسبه و اضافه می‌گردد.

این سه ویژگی با نام‌های PacketSumLen و SumTimeSinceFirstFrame و SumTimeSincePreviousFrame در مجموع ویژگی‌های مورد مطالعه قرار می‌گیرد.

در ستون آخر جدول ویژگی‌ها نیز وضعیت منبع اعم از آلوده یا پاک بودن قرار می‌گیرد.(Isworm)

در مرحله آخر با نگاه به خاصیت علت و معلولی کرم که پیشتر به آن اشاره شد تمامی خصیصه‌های فوق را برای سطح دو و سطح سه محاسبه می‌نماییم.

سطح دو بدین معناست که اگر میزبان A روی درگاه a با میزبان B ارتباط برقرار نماید و آنگاه میزبان B روی درگاه a با میزبان C ارتباط برقرار نماید، میزبان C در سطح دوم ارتباط میزبان A قرار خواهد گرفت. در این مرحله برای هر سطح به کمک مجموع‌گیری مقادیر ویژگی‌های مرتبط با آن سطح را ایجاد می‌نماییم.

این سطح‌بندی با تمرکز بر واقعیت تمایل به انتشار آلودگی توسط میزبان آلوده به سطوح دیگر حاصل شده است.

بنابراین، نه بردار ویژگی در سه سطح برای هر میزبان و درگاه وی ایجاد می‌گردد که به‌عنوان ورودی خوشه‌بندی و دسته‌بندی در داده‌کاوی مورد مطالعه قرار خواهد گرفت.

مجموعه این ویژگی‌ها در یک نگاه مطابق جدول (۱) می‌باشد

جدول (۱): مجموعه ویژگی‌های احصا شده

ردیف	نام لاتین	نام فارسی	شرح ویژگی
۱	Connection Status	وضعیت برقراری اتصال	در صورت موفقیت‌آمیز بودن برقراری اتصال مقدار آن یک می‌باشد
۲	One	تعداد رکوردهای تک اتصالی میزبان	تعداد اتصالاتی که بین میزبان و مقصد فقط یک‌بار برقرار شده است.
۳	Plus01	تعداد رکوردهای یک تا ۵ اتصالی میزبان	تعداد اتصالاتی که بین میزبان و مقصد بیش از یک‌بار و کمتر از ۵ بار برقرار شده است
۴	Plus05	تعداد رکوردهای پنج تا۱۰ اتصالی میزبان	تعداد اتصالاتی که بین میزبان و مقصد بیش از ۵ بار و کمتر از۱۰ بار برقرار شده است.
۵	Plus10	تعداد رکوردهای بیش از ۱۰ اتصالی میزبان	تعداد اتصالاتی که بین میزبان و مقصد بیش از۱۰ بار برقرار شده است.
۶	PacketSumLen	مجموع اندازه بسته‌های تبادل شده	مجموع اندازه بسته‌های مبادله شده توسط میزبان روی درگاه X برای هر وضعیت اتصال
۷	SumTime Since FirstFrame	مجموع فاصله زمانی بسته‌های تبادل	مجموع فاصله زمان بسته‌ها، نسبت به زمان آغازین اتصال به‌ازای هر بسته
۸	SumTime Since PreviousFrame	مجموع فاصله‌های زمانی بسته‌ها	مجموع فاصله‌های زمانی بسته‌های متوالی برای هر میزبان
۹	IsWorm	وضعیت پاک یا آلوده بودن	در صورت صفر بودن بسته پاک و در صورت یک بودن آلوده (ناشی از ترافیک کرم) می‌باشد

۴-۲. مدل‌سازی

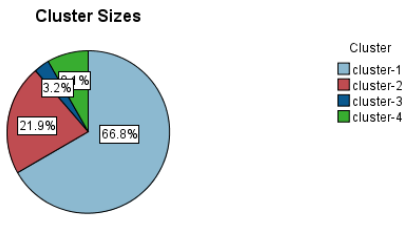
مدل‌سازی در این تحقیق شامل دو بخش می‌باشد. در بخش اول داده‌های مربوط به رکوردهایی که کرم نبوده‌اند وارد روش‌های خوشه‌بندی شده و پس از پیدا کردن بهترین روش خوشه‌بندی با رکوردهایی که کرم بوده ترکیب شدند و مجموعه داده‌های آموزشی جدیدی را تولید کرده‌اند. در بخش دوم این مجموعه داده‌ها به‌طور جداگانه وارد روش‌های رده‌بندی می‌گردند. هنگام ورود رکورد جدید مشخص می‌شود که متعلق به کدام خوشه بوده و با توجه به خوشه آن مدل مربوط به آن خروجی مورد استفاده قرار می‌گیرد.

۴-۲-۱. متوازن‌سازی مقادیر خروجی مدل‌سازی

پس از انجام نمونه‌گیری طی بررسی به‌عمل آمده، حدود ۹۰ درصد ترافیک مرتبط با منابع پاک و ۱۰ درصد آن مرتبط با منابع آلوده می‌باشد بدیهی است که در صورت عدم ایجاد توازن بر روی داده‌ها مدل با مشاهده داده‌های پاک اقدام به تشخیص آنها نموده و از داده‌های آلوده صرف‌نظر می‌نماید. برای جلوگیری از این پدیده لازم است تا مجموعه داده‌ها متوازن گردد. این متوازن‌سازی باید طوری انجام شود تا بدون تصویرسازی غلط از داده‌ها برای مدل بتوان بر مشکل موجود غلبه پیدا نمود برای این منظور به دو صورت تصادفی و استفاده از خوشه‌بندی داده‌های پاک عملیات متوازن‌سازی را انجام داده و نتایج را در مرحله مدل‌سازی مورد مقایسه قرار می‌دهیم.

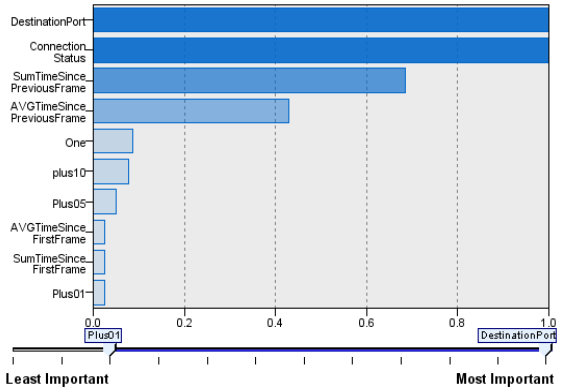
۴-۲-۲. خوشه‌بندی

به‌منظور خوشه‌بندی ابتدا رکوردهایی که کرم نبودند جدا شده و وارد روش‌های خوشه‌بندی شدند. در این تحقیق به‌منظور خوشه‌بندی از روش خوشه‌بندی کی- میانگین استفاده شده است. در روش کی- میانگین تعداد خوشه‌ها توسط کاربر تعیین می‌گردد. بدین ترتیب به‌منظور یافتن خوشه‌ها با استفاده از روش کی- میانگین از تعداد ۲ خوشه تا ۱۵ خوشه جستجو شد و از طرفی دیگر در روش دو مرحله‌ای فقط حداقل و حداکثر خوشه‌ها مشخص شدند. با تعداد خوشه‌های مختلف تعداد خوشه بهینه ۴ خوشه به‌دست آمد توزیع نمونه‌های موجود در هر خوشه مطابق با شکل (۲) به‌دست آمده است.



شکل (۲): توزیع نمونه‌های موجود در هر خوشه

همچنین در شکل (۳) متغیرهایی که بیشترین تأثیر را بر روی خوشه‌بندی داشته‌اند، به همراه میانگین مقادیر این متغیرها در هر خوشه و همچنین تعداد رکوردهای قرار گرفته در هر خوشه ذکر شده است.



شکل (۳): اهمیت متغیرها در خوشه‌بندی

۴-۲-۳. رده‌بندی

هدف این بخش به‌کارگیری روش‌های رده‌بندی در تشخیص میزبان‌های آلوده و پاک می‌باشد برای این منظور سه روش ماشین بردار پشتیبان، درخت تصمیم و شبکه‌های عصبی مصنوعی به‌عنوان روش‌های رده‌بندی مورد استفاده قرار دادیم نتایج را یک‌بار در وضعیت عدم استفاده از خوشه‌بندی و به‌کارگیری متوازن‌سازی تصادفی و بار دیگر در حالت به‌کارگیری خوشه‌بندی به‌دست آورده و آنها را ارزیابی نمودیم.

در مدل‌سازی نیز به‌منظور یادگیری و ارزیابی از روش ۸۰-۲۰ استفاده نمودیم. بدین ترتیب در هر دسته از مجموعه داده‌های ساخته‌شده از ۸۰٪ رکوردها به‌منظور آموزش مدل و از ۲۰٪ رکوردها به‌منظور آزمایش مدل استفاده کردیم.

در روش‌های داده‌کاوی برای ارزیابی کلی معمولاً از معیار صحت[1](مجموعه مواردی که به‌درستی آلوده تشخیص داده شده و مجموعه مواردی که به‌درستی پاک تشخیص داده شده نسبت به کل مجموعه موارد) استفاده می‌گردد اما در مواردی که فراوانی حالات مطلوب در بین کل حالات کم است این معیار به تنهایی نمی‌تواند دقت مدل را توصیف نماید برای همین منظور از معیارهای دقت[2] (نسبت تعداد مواردی که به‌درستی آلوده تشخیص داده شده[3] به کل مواردی که درست آلوده تشخیص داده شده و یا غلط آلوده تشخیص داده شده[4]) و معیار بازخوانی[5]

[1] Accuracy
[2] Precision
[3] True Positive
[4] False Negative
[5] Recall

(نسبت تعداد مواردی که به‌درستی آلوده تشخیص داده شده به کل مواردی که درست آلوده تشخیص داده شده و یا به غلط پاک تشخیص داده شده است[1]) و ترکیب این دو معیار در قالب سنجش معیار اِف[2] (دوبرابر حاصل‌ضرب معیار سنجش در معیار بازخوانی نسبت به حاصل‌جمع این دو معیار) استفاده می‌نماییم.

- **به‌کارگیری ماشین بردار پشتیبان به کمک متوازن‌سازی تصادفی**

در این وضعیت بدون انجام خوشه‌بندی ذکرشده بر روی ترافیک ورودی میزان آلوده یا پاک بودن خروجی بسته‌های اطلاعاتی به کمک بردار پشتیبان مورد ارزیابی واقع شده است. همان‌طور که ملاحظه می‌نمایید از مجموع ۱۴۷۲ بسته مورد ارزیابی تعداد ۱۲۱۰ مورد به‌درستی متعلق به منبع پاک یا آلوده شناسایی شده که صحت مدل برابر ٪۸۲/۲۰ می‌گردد.

جدول (۲): به‌کارگیری ماشین بردار پشتیبان بدون اتکا به خوشه‌بندی

مشاهده‌شده	پیش‌بینی‌شده		
		آلوده	پاک
	آلوده	۵۶۹	۱۱۱
	پاک	۱۵۱	۶۴۱

- **به‌کارگیری شبکه‌های عصبی MLP به کمک متوازن‌سازی تصادفی**

در این وضعیت نیز مانند روش قبل فارغ از عملیات خوشه‌بندی ارزیابی خود را انجام می‌دهیم، همان‌طور که ملاحظه می‌نمایید از مجموع ۱۵۲۲ بسته مورد ارزیابی تعداد ۱۳۵۹ مورد به‌درستی متعلق به منبع پاک یا آلوده شناسایی شده که صحت مدل برابر ٪۸۹/۲۹ می‌گردد.

همچنین لازم به ذکر است که مدل ما به‌طور متناظر ٪۱۷/۸۰ موارد را به اشتباه آلوده یا پاک تشخیص داده است.

این مدل قادر است تا با دقت ٪۷۹/۰۲ و با بازخوانی ٪۸۳/۶۷ و معیار اِف ۸۱/۲۷ موارد آلوده را از پاک مشخص سازد.

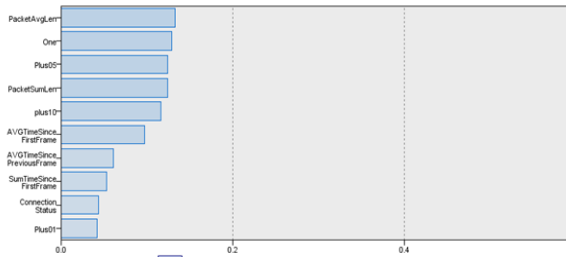
جدول (۳): به‌کارگیری شبکه‌های عصبی بدون اتکا به خوشه‌بندی

مشاهده‌شده	پیش‌بینی‌شده		
		آلوده	پاک
	آلوده	۵۷۳	۱۰۷
	پاک	۵۶	۷۸۶

همچنین لازم به ذکر است که به‌طور متناظر مدل ٪۱۰/۷۱ موارد را به اشتباه آلوده یا پاک تشخیص داده است. این مدل قادر است تا با دقت ٪۹۱/۰۹ و با بازخوانی ٪۸۴/۲۶ و معیار اِف ۸۷/۵۴ موارد آلوده را از پاک مشخص سازد.

شکل (۴) متغیرهای تأثیرگذار در هر مجموعه داده را از دید شبکه‌های عصبی نشان می‌دهد همان‌گونه که ملاحظه می‌نمایید سه دسته متغیر بیشترین تأثیر را در ایجاد مدل ایفا نمودند.

دسته اول متغیرهای مرتبط با اندازه بسته‌ها می‌باشند (نظیر PacketAvgLen) که الگویی از شباهت اندازه بسته‌ها را در بسته‌های ناشی از ترافیک آلوده و پاک دنبال می‌نمایند. دسته دوم متغیرهای مرتبط با تعداد اتصالات موفق یا ناموفق هستند (نظیر PlusOne) که با رفتار پویش سر و کار دارند و دسته سوم نیز متغیرهای در ارتباط با مدت زمان برقراری اتصال می‌باشند که الگویی مشابه را در بسته‌های آلوده دنبال می‌کنند. (نظیر AVGTimeSincePreviousFrame)



شکل (۴): متغیرهای تأثیرگذار در هر مجموعه داده از دید شبکه‌های عصبی MLP

- **به‌کارگیری درخت تصمیم C5 به کمک متوازن‌سازی تصادفی**

همان‌طور که ملاحظه می‌نمایید از مجموع ۱۵۴۲ بسته مورد ارزیابی تعداد ۱۴۵۷ مورد به‌درستی متعلق به منبع پاک یا آلوده شناسایی شده که صحت مدل برابر ٪۹۴/۴۹ می‌گردد.

جدول (۴): به‌کارگیری درخت تصمیم C5 بدون اتکا به خوشه‌بندی

مشاهده‌شده	پیش‌بینی‌شده		
		آلوده	پاک
	آلوده	۶۴۴	۳۶
	پاک	۴۹	۸۱۳

همچنین لازم به ذکر است که به‌طور متناظر مدل ٪۵/۵۱ موارد را به اشتباه آلوده یا پاک تشخیص داده است. این مدل قادر است تا با دقت ٪۹۲/۹۲ و با بازخوانی ٪۹۴/۷۰ و معیار اِف ۹۳/۸۰ موارد آلوده را از پاک مشخص سازد.

[1] False Negative
[2] F-measure

• رده‌بندی مبتنی بر به‌کارگیری خوشه‌بندی بر روی داده‌های پاک

به‌طور کلی در مجموعه مدل‌های ارائه‌شده تکیه بر خوشه‌بندی می‌تواند از دو دیدگاه تأثیری ملموس در نتایج به‌دست آمده ایجاد نماید. همان‌گونه که پیشتر اشاره شد یکی از اساسی‌ترین مشکلات در ارائه مدل مناسب متوازن‌سازی داده‌های به‌کارگیری‌شده می‌باشد هرچند به‌کارگیری روش‌های متوازن‌سازی تصادفی به‌عنوان در دسترس‌ترین راه امکان‌پذیر می‌باشد ولی هدفمند نمودن این متوازن‌سازی به کمک روش‌های خوشه‌بندی معمولاً می‌تواند نتایج بهتری را به ارمغان آورد از طرف دیگر تحلیل رفتاری بسته‌های پاک نیز مسئله‌ای است که نباید تحت هیچ عنوان از مدل‌سازی کنار گذاشته شود و برای دستیابی به این مهم خوشه‌بندی می‌تواند نقش مؤثری را ایفا نماید برای این منظور ما عملیات خوشه‌بندی را بر روی بسته‌های دارای برچسب پاک انجام داده و چهار خوشه به‌دست آوردیم از طرف دیگر این چهار خوشه را به‌صورت تک تک با خوشه بسته‌های آلوده مدل‌سازی نموده تا مسئله متوازن‌سازی پوشش داده شود. هنگام ورود یک بسته جدید با توجه به خوشه‌ای که در آن قرار می‌گیرد تصمیم می‌گیریم کدام مدل را بر روی آن اعمال نماییم بنابراین، به‌جای روش تصادفی متوازن کردن با به‌کارگیری روش الگوهای مشابه می‌توانیم صحت مدل را افزایش دهیم.

از آنجا که بهترین نتایج در درخت تصمیم C5 حاصل شده خوشه‌بندی مذکور را برای این مدل استفاده می‌نماییم. نتایج مطابق با جدول (۵) حاصل می‌گردد.

جدول (۵): به‌کارگیری درخت تصمیم C5 بر روی خوشه‌های تشکیل‌شده

مشاهده‌شده			پیش‌بینی‌شده				
			آلوده	پاک	صحت (٪)	دقت (٪)	بازخوانی (٪)
خوشه یک (۱۰۰۴۵)	آلوده	۵۷۶	۱۰۴	۹۸/۹۴	۹۹/۴۸	۸۴/۷۰
	پاک	۳	۹۳۶۲			
خوشه دو (۲۱۴۲)	آلوده	۶۶۸	۱۲	۹۹/۳۹	۹۹/۸۵	۹۸/۲۳
	پاک	۱	۱۴۶۱			
خوشه سه (۷۶۴)	آلوده	۶۷۶	۴	۹۹/۳۴	۹۹/۸۵	۹۹/۴۱
	پاک	۱	۸۳			
خوشه چهار (۳۴۱۶)	آلوده	۴۷۳	۲۷	۹۸/۷۱	۹۶/۵۳	۹۴/۶۰
	پاک	۱۷	۲۸۹۹			

به‌منظور ارزیابی پیشرفت حاصل‌شده در این حالت از میانگین وزنی چهار خوشه استفاده می‌نماییم.

شاخص صحت برابر ۹۹/۰۷، دقت برابر ۹۹/۰۳، بازخوانی برابر ۸۹/۳۲ و معیار اِف برابر ۹۳/۹۲ می‌گردد.

با درنظر گرفتن نتایج مرحله قبل همان‌گونه که ملاحظه می‌نمایید شاخص صحت از ۹۴/۴۹ به ۹۹/۰۷، شاخص بازخوانی از ۹۴/۷۰ به ۸۹/۳۲ ، شاخص دقت از ۹۲/۹۲ به ۹۹/۰۳ و معیار اِف از ۹۳/۸ به ۹۳/۹۲ تغییر می‌یابد که نشانگر موفق عمل نمودن ترکیب ذکرشده بر روی داده‌های مورد آزمایش می‌باشد.

۴-۳. مقایسه با روش‌های مشابه

در پژوهش انجام‌شده توسط پراس و همکارانش [۲] آنها توانستند با دقت ۷۰ درصد به کمک شبکه‌های عصبی از ویژگی‌های طول، اندازه بسته، زمان ارسال و آدرس‌های مقصد بسته‌های ترافیک بدافزارها را از ترافیک عادی تشخیص دهند آنها با تکیه بر الگوی رشته‌های آدرس دامنه نتایج خود را تا حدی بهبود بخشیدند و با دقت ۹۰ و بازخوانی ۸۰ درصد بر روی مجموعه تستی خود نتایج بهتری را ارائه نمودند.

ما در روش خود با ایجاد چندین بهبود توانستیم نتایج بهتری را در اکثر حالات ایجاد نماییم. همراه کردن مدل با وضعیت اتصالات در کشف پویش، به‌کارگیری ایده گسترش‌های علت و معلولی و سطح‌بندی ارتباطات، استفاده از مدت زمان بقای اتصال در انتخاب مجموعه ویژگی‌ها همچنین به‌کارگیری خوشه‌بندی بر روی ترافیک‌های پاک و استفاده از الگوهای هنجار، متوازن‌سازی مبتنی بر خوشه‌بندی و در نهایت به‌کارگیری درخت تصمیم بر روی داده‌ها سبب شد تا علاوه بر افزایش دقت در کشف بسته‌های آلوده وابستگی مدل از برقراری ارتباط با دامنه‌های آلوده نیز برداشته شود.

همچنین مدل پراس با شرط انتشار بسته‌ها به کمک پروتکل HTTPS و وجود URL مشخص در بسته‌ها توانست به این نتایج دست یابد و این در حالی است که در مدل ما هیچ قیدی برای پروتکل انتشار کرم وجود نداشته و صرفاً از سرآیند ترافیک TCP عمل تشخیص انجام می‌پذیرد.

۴-۴. تشخیص کرم از میزبان آلوده

بنابر آنچه ذکر شد با توجه به مدل‌سازی انجام‌شده در این بخش مشخص است که روش رده‌بندی درخت‌های تصمیم نتایج بهتری را در تشخیص بسته‌های آلوده از پاک حاصل می‌نماید و همچنین به‌منظور بالا بردن صحت مدل، استفاده از خوشه‌بندی و به‌دست آوردن الگوهای رفتارهای سالمی که کرم نیستند نتایج بهتری را در تشخیص کرم به ارمغان خواهد آورد.

در پایان لازم است یادآور شویم که بررسی و یافتن میزبان‌های آلوده گام اصلی تشخیص کرم در شبکه می‌باشد زیرا

به‌محض یافتن میزبان آلوده تمامی تراکنش فایل‌ها با پروتکل‌های شبکه و سوابق اتصالات آنها بررسی می‌گردد[1] و فایل‌ها و توابع کتابخانه‌ای اجرایی بدافزارها استخراج می‌شود. این فایل‌ها و توابع برای بررسی بیشتر و عملیات تحلیل به آزمایشگاه تحلیل بدافزار سپرده می‌شود. مراحل مواجهه با بدافزار و جرمیابی کار خارج از حیطه تحلیل بدافزار بوده و در مجال این مقاله نمی‌گنجد.

۵. نتیجه‌گیری

۵-۱. جمع‌بندی

آنچه در این مقاله به آن پرداخته شده روشی برای یافتن کرم‌های نهان می‌باشد. کرم‌هایی که هنوز برای آنها در آنتی‌ویروس‌ها امضایی وجود نداشته، در شبکه محتوای خود را رمز نموده و می‌توانند از یک میزبان به میزبان دیگر شکل خود را تغییر دهند.

در بخش اول این مقاله با مطالعه ویژگی‌های کرم و انتخاب قابلیت خودتکثیری و استفاده از آسیب‌پذیری‌های سیستم به‌عنوان خصیصه‌های اصلی کار خود را آغاز نمودیم. در بخش دوم ویژگی‌های تشخیص‌دهنده مناسب را بررسی نمودیم و در بخش سوم انواع روش‌های مقابله با کرم اعم از میزبان محور در نقاط پایانی شبکه، شبکه محور در روی دیواره آتش و نقاط میانی شبکه و هانی‌پات محور در بخشی ایزوله از شبکه را ارزیابی نموده و تمرکز خود را بر روی روش‌های شبکه محور قرار دادیم. با درنظرگرفتن ویژگی‌های کرم در می‌یابیم کشف میزبان‌های پویشگر می‌تواند گام بزرگی در کشف میزبان‌های آلوده و در نهایت کشف کرم باشد برای این منظور در بخش چهارم مدلی برای کشف رفتار پویش و ویژگی‌های انتشار کرم در شبکه ارائه نمودیم.

در مدل ارائه‌شده از روش‌های داده‌کاوی و مدل کریسپ برای کشف ترافیک آلوده از پاک بهره گرفتیم. بنابراین، ابتدا ویژگی‌های کلیدی قابل بررسی را از ترافیک آزمایشی به‌دست‌آمده متعلق به آزمایشگاه ICIS استخراج نمودیم. برای این منظور تمرکز خود را بیشتر بر روی روش‌های کشف کرم در شبکه که پیشتر به آنها اشاره شد قرار دادیم و با تلفیقی از ویژگی‌های کلیدی آنها ویژگی‌های خود را ایجاد نمودیم. یکی از مهم‌ترین نوآوری‌های ارائه‌شده استخراج ویژگی‌هایی نظیر تعداد اتصالات موفق و ناموفق هر میزبان، طول و زمان ارسال بسته‌ها، الگوی انتخاب آدرس‌ها و درگاه‌های مقصد می‌باشد که از دل دیگر صفات مورد بررسی در داده‌های آزمایشگاه استنتاج گردید. برای ارزیابی نیز شاخص‌های دقت و بازخوانی را به مجموعه شاخص صحت روش خود اضافه نمودیم تا بتوانیم ارزیابی دقیقی در شرایط توزیع‌شدگی ترافیک آلوده کم در بین ترافیک پاک زیاد ایجاد نماییم.

در بخش پنجم نیز پیاده‌سازی خود را به کمک نرم‌افزار کلمنتاین انجام دادیم. با توجه به عدم توازن داده‌های آلوده و پاک لازم بود تا این توازن را ایجاد می‌کردیم برای این منظور یک‌بار به کمک روش‌های تصادفی توازن را صورت بخشیده و بهترین نتایج روش‌های رده‌بندی را که شامل ماشین بردار پشتیبان، شبکه‌های عصبی و درخت تصمیم C5 بود مورد ارزیابی قرار دادیم. برای یادگیری از ۸۰ درصد ترافیک و برای آزمون از ۲۰ درصد آن استفاده نمودیم. در مجموع بردار پشتیبان با صحت ٪۸۲/۲، دقت ٪۷۹/۰۲ و با بازخوانی ٪۸۳/۶۷ ، شبکه‌های عصبی با صحت ٪۸۹/۲۹ ، دقت ٪۹۷/۰۹ و با بازخوانی ٪۸۴/۲۶ ، درخت تصمیم C5 با صحت ٪۹۴/۴۹ ، دقت ٪۹۲/۹۲ و با بازخوانی ٪۹۴/۷۰ توانستند بهترین نتایج را در کشف بسته‌های آلوده از پاک ارائه نمایند.

مدل را تکامل بیشتری بخشیدیم و در نوآوری خود به‌جای متوازن‌سازی تصادفی از خوشه‌بندی در متوازن‌سازی استفاده کردیم. این روش باعث شد تا علاوه بر هدفمند نمودن متوازن‌سازی از ترافیک پاک استفاده مؤثرتری در عملیات تشخیص بنماییم. نتایج را بر روی رده‌بندی C5 اعمال نمودیم و همان‌گونه که انتظار می‌رفت در برخی دسته‌ها تا ۹۹ درصد دقت، صحت و بازخوانی را بهبود بخشیدیم.

۵-۲. پیشنهاد برای کارهای آینده

بدافزارهای مانای پیشرفته[2] نوعی جدید از بدافزارهای روزصفرم می‌باشد که از جمله ویژگی‌های آنها پاک کردن فایل مولد بعد از انتشار روی هدف می‌باشد. این بدافزارها در حافظه اصلی مقیم شده و معمولاً از رایانه‌های کارگزار انتشار خود را صورت می‌بخشند. برای این منظور لازم است تا با استفاده از روش‌های تصویربرداری از حافظه اصلی عملیات داده‌کاوی بر روی این حافظه صورت پذیرفته و نشر بدافزار تشخیص داده شود. پیشنهاد می‌گردد ضمن بررسی فراخوانی‌های شبکه موجود در حافظه اصلی به کمک داده‌کاوی مدلی ارائه گردد تا انتشار بدافزار از روی کارگزار مربوطه تشخیص داده شود.

[1] دستوراتی نظیر Netstat –ao در Cmd مجموعه تمام پروسس‌هایی که سعی در برقراری ارتباط با درگاه‌ها و آدرس‌های شبکه را داشته‌اند ارائه می‌دهد. بنابر این با مراجعه به سیستم‌های آلوده می‌توان فایل‌های مخرب را احصا نمود. البته باید توجه داشت که مراجعه دیرتر از موعد ممکن است نتیجه بخش نبوده و کرم بخش مخرب خود را از بین برده باشد.

[2] Advanced Persistent Threat

۶. مراجع

[1] K. Bartos and M. Sofka, ‘‘Robust representation for domain adaptation,’’ In European Conference on Machine Learning and Knowledge Discovery in Databases, 2015.

[2] P. Prasse, G. Gruben, L. Machlika, T. Pevny, M. Sofka, and T. Scheffer, ‘‘Malware Detection by HTTPS Traffic Analysis,’’ Universität Potsdam, no. computer security, machine learning, p. 10, 2017.

[3] ICSI, ‘‘International Computer Science Institute,’’ Berkeley University of California, [Online]. Available: http://www.icir.org/enterprise-tracing/download.html.

.

*Iran University of Science and Technology

(Received: 10/09/2018, Accepted: 13/10/2018)

Complex malwares which infiltrate systems in a country's critical infrastructure with the purpose of destruction or espionage are major threats in cyber space. What is presented in this article is a smart solution to discover *zero day worms which can be polymorphic and encrypted and their nature is still unknown to defense tools.To do this, we first outlined our desirable detector and then presented a solution based on data mining methods for detecting malicious extensions with the emphasis on worm's scanning feature, communication model of the infected hosts and the packets' headers transmitted across the network.By clustering clean data, and using clean and contaminated data classifications, experimental samples and the C5 decision tree, we managed to present the best model with an accuracy of 94.49%, precision of 92.92%, and a recall of 94.70% in identifying infected packages from the clean ones. Finally, we also showed that the use of clustering in the patterns of clean hosts' traffic could reach better results in identifying infected traffic.*

Worm, Detector, Zero day, Scan, Data mining

* Corresponding Author Email: b_minaei@iust.ac.ir