به نام دانش و خرد

واحد بین‌الملل

پایان‌نامه کارشناسی ارشد در رشته مهندسی کامپیوتر
(هوش مصنوعی)

ارائه راهکاری نوین برای شناسایی بدافزارهای ناشناخته بر اساس توالی کدهای عملیاتی اسمبلی

به کوشش

هاشم هاشمی

استاد راهنما:

دکتر شهرام جعفری

۱۹ اسفندماه ۱۳۹۲

به نام خدا

اظهارنامه

این‌جانب هاشم هاشمی(۹۰۸۸۸۴) دانشجوی رشته‌ی مهندسی کامپیوتر گرایش هوش مصنوعی واحد بین‌الملل دانشگاه شیراز اظهار می‌کنم که این پایان‌نامه حاصل پژوهش خودم بوده و در جاهایی که از منابع دیگران استفاده کرده‌ام، نشانی دقیق و مشخصات کامل آن را نوشته‌ام. همچنین اظهار می‌کنم که تحقیق و موضوع پایان‌نامه‌ام تکراری نیست و تعهد می‌نمایم که بدون مجوز دانشگاه دستاوردهای آن را منتشر ننموده و یا در اختیار غیر قرار ندهم. کلیه حقوق این اثر مطابق با آیین‌نامه مالکیت فکری و معنوی متعلق به دانشگاه شیراز است.

نام و نام خانوادگی: هاشم هاشمی

تاریخ و امضاء: ۹۲/۱۲/۱۹

به نام خدا

ارائه راهکاری نوین برای شناسایی بدافزارهای ناشناخته بر اساس توالی کدهای عملیاتی اسمبلی

به‌وسیله :

هاشم هاشمی

پایان‌نامه

ارائه‌شده به تحصیلات تکمیلی دانشگاه به‌عنوان بخشی

از فعالیت‌های تحصیلی لازم برای اخذ درجه کارشناسی ارشد

در رشته‌ی:

مهندسی کامپیوتر – هوش مصنوعی

از دانشگاه شیراز

شیراز

جمهوری اسلامی ایران

ارزیابی‌شده توسط کمیته پایان‌نامه با درجه :

دکتر شهرام جعفری، استادیار بخش مهندسی و علوم کامپیوتر و فن‌آوری اطلاعات(رئیس کمیته).....

دکتر علی حمزه، استادیار بخش مهندسی و علوم کامپیوتر و فن‌آوری اطلاعات.....

دکتر ستار هاشمی، استادیار بخش مهندسی و علوم کامپیوتر و فن‌آوری اطلاعات.....

اسفندماه ۱۳۹۲

تقدیم به پدر و مادر عزیزم

خدای را بسی شاکرم که از روی کرم، پدر و مادری فداکار نصیبم ساخته تا در سایه درخت پر بار وجودشان بیاسایم و از ریشه آنها شاخ و برگ گیرم و از سایه وجودشان در راه کسب علم و دانش تلاش نمایم.

والدینی که بودنشان تاج افتخاری است بر سرم و نامشان دلیلی است بر بودنم چرا که این دو وجود پس از پروردگار مایه هستی ام بوده اند دستم را گرفتند و راه رفتن را در این وادی زندگی پر از فراز و نشیب آموختند.

آموزگارانی که برایم زندگی؛ بودن و انسان بودن را معنا کردند

حال این برگ سبزی است تحفه درویش تقدیم آنان. . . .

سپاسگزاری

اکنون که این پایان‌نامه به پایان رسیده است بر خود لازم می‌دانم تا از زحمات بی‌دریغ اساتید بزرگوارم جناب آقای دکتر شهرام جعفری و جناب آقای دکتر علی حمزه، که از آغاز تا پایان کار باراهنمایی‌های ارزشمند خود زمینه‌ساز پیشرفت پایان‌نامه شدند و در این راه زحمات فراوانی را بر دوش گرفتند، نهایت سپاس و قدردانی را داشته باشم.

همچنین از اساتید بزرگوار، جناب آقای دکتر ستار هاشمی و جناب آقای دکتر محمدحسین شیخی که به‌عنوان استاد داور و نماینده تحصیلات تکمیلی در ارائه این پژوهش بنده را همراهی کردند سپاسگزارم.

چکیده

ارائه راهکاری نوین برای شناسایی بدافزارهای ناشناخته بر اساس توالی کدهای عملیاتی اسمبلی

به کوشش

هاشم هاشمی

بدافزارها[1]، برنامه‌های مخربی هستند که بدون اجازه کاربر، به سیستم کامپیوتری آنها نفوذ کرده و آن‌ها را آلوده می‌کنند و دست به اعمال خرابکارانه می‌زنند. این برنامه‌های مخرب، به طور روزافزون در حال گسترش هستند و باگذشت زمان شاهد تکامل و پیشرفت‌های چشم‌گیری بوده‌اند، به‌طوری که روش‌های سنتی شناسایی بدافزار مانند روش‌های مبتنی بر امضا قادر به شناسایی انواع نوظهور بدافزارها نمی‌باشند، همچنین تولیدکنندگان بدافزارها از استراتژی‌های متنوع مبهم سازی و اختفا به‌منظور فرار از شناسایی شدن توسط روش‌های سنتی استفاده می‌کنند و این امر موجب بروز یک بحران امنیتی در عصر حاضر شده است. در مقابل، محققان این حوزه می‌بایست تدبیری برای مقابله با این بحران باندیشند و با ارائه روش‌های کارآمد در مقابل این بحران بزرگ ایستادگی کنند. لذا، لزوم ارائه روش‌های نوین که قادر به شناسایی بدافزارهای نوظهور باشند، بیش‌ازپیش محسوس است. در همین راستا، در این پایان‌نامه، یک روش نوین شناسایی بدافزار بر پایه توالی کدهای عملیاتی زبان اسمبلی موجود در یک فایل اجرایی، ارائه شده است. در این روش با استفاده از ابزارهای یادگیری ماشین سعی شده از توزیع کدهای عملیاتی موجود در هر دو کلاس بدافزار و فایل‌های خوش‌خیم[2] و همچنین توزیع درون کلاسی کدهای عملیاتی موجود در هر یک از کلاس‌ها بهره برده شود و با استفاده از این توزیع به تفکیک بدافزارها از فایل‌های سالم پرداخته شود. از مزایای این روش می‌توان به نرخ شناسایی بالا، پیچیدگی محاسباتی قابل‌قبول، نرخ مثبت کاذب پایین و همچنین قابلیت توسعه آن اشاره کرد که این قابلیت توسعه، این امکان را فراهم می‌آورد که با استفاده از روش ارائه‌شده در این پایان‌نامه، قادر به دسته‌بندی بدافزارها بر اساس نوع (خانواده) آنها باشیم.

واژگان کلیدی: شناسایی بدافزارهای ناشناس، توالی کدهای عملیاتی، یادگیری ماشین

---

[1] Malware

[2] Benign

فهرست مطالب

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

فصل نخست:

مقدمه

فصل ۱- مقدمه

۱-۱- شرح مسئله:

در عصر ارتباطات و دنیای دیجیتال، امنیت سامانه‌های کامپیوتری به یکی از بحث‌برانگیزترین موضوعات امنیتی بدل شده است. این امنیت از دیدگاه‌های مختلفی قابل بررسی است که یکی از مهم‌ترین آنها امنیت اطلاعات و مقابله با انواع بدافزارها[۱] می‌باشد.

بدافزارها، برنامه‌های بدخواهی هستند که بدون اجازه کاربر به سیستم‌های کامپیوتری آنها نفوذ کرده و آن‌ها را آلوده می‌کنند و دست به اعمال خرابکارانه‌ای از قبیل سرقت اطلاعات، دسترسی غیرمجاز، تخریب اطلاعات و غیره می‌زنند. بدافزارها بر اساس عملکرد، به گونه‌های مختلفی تقسیم می‌شوند که مهم‌ترین آنها عبارت‌اند از ویروس‌ها[۲]، تروجان‌ها[۳]، درب‌های پشتی[۴]، کرم‌ها[۵]، جاسوس‌افزارها[۶] ، روت کیت‌ها[۷]، بات‌نت‌ها[۸]، ترس‌افزارها[۹] و تبلیغ افزارها[۱۰]. هرکدام از

---

[۱] Malware
[۲] Viruses
[۳] Trojans
[۴] Back Doors
[۵] Worms
[۶] Spywares
[۷] Rootkits
[۸] Botnets
[۹] Scarewares
[۱۰] Spywares

انواع بدافزارها باهدف خاصی به وجود آمده و بسته به نوع خود، عملیات خرابکارانه خاص خود را به انجام می‌رسانند.

تعداد بدافزارها هر روزه به طور چشمگیری در حال افزایش است، همچنین تولیدکنندگان بدافزارها از تکنیک‌های پیشرفته برنامه‌نویسی و روش‌های نوین اختفا[۱] از جمله تکنیک‌های مبهم سازی[۲] ، رمزگذاری[۳] ، استراتژی‌های چندشکلی[۴] و استراتژی‌های دگرگون شونده[۵] و غیره در توسعه بدافزارهای خود استفاده می‌کنند و همین امر، عملیات شناسایی بدافزارهای نوظهور را به امری بسیار دشوار و چالش‌برانگیز بدل کرده است. این پیشرفت‌ها در توسعه بدافزارها از یک سو، همچنین اهمیت امنیت سیستم‌های کامپیوتری از سوی دیگر، مقابله با این تهدید بزرگ را به یکی از مباحث به‌روز در حوزه امنیت سیستم‌های کامپیوتری تبدیل کرده است، لذا تحقیقات وسیعی به‌موازات پیشرفت‌های صورت گرفته از سوی تولیدکنندگان بدافزارها، توسط محققان این حوزه، در زمینه شناسایی بدافزارها در حال انجام است. این تحقیقات باعث به وجود آمدن روش‌های نوینی در شناسایی بدافزارها شده است که این روش‌ها را می‌توان به سه گروه مختلف دسته‌بندی کرد که در ادامه، آنها را نام‌برده و به شرح هر یک از آنها می‌پردازیم.

۱-۲- روش‌های شناسایی بدافزار[۶]:

شناسایی بدافزار شاخه‌ای از امنیت سیستم‌های کامپیوتری است که سعی در تشخیص، تحلیل و مقابله با انواع بدافزارها دارد. در سال‌های اخیر روش‌های مختلفی برای شناسایی بدافزارها مورد استفاده قرار گرفته است که این روش‌ها را می‌توان به سه گروه کلی روش‌های شناسایی مبتنی

---

۱ Concealment Strategies
۲ Obfuscation
۳ Encryption
۴ Polymorphism
۵ Metamorphism
۶ Malware detection techniques

بر امضا[1]، روش‌های شناسایی ناهنجاری[2] و روش‌های مکاشفه‌ای[3] شناسایی بدافزار تقسیم‌بندی کرد که در ادامه به بررسی هر یک می‌پردازیم.

۱-۲-۱- روش‌های مبتنی بر امضا:

روش‌های مبتنی بر امضا، روش‌های سنتی‌ای است که بر پایه پایگاه داده‌ای از امضاهای استخراج‌شده از بدافزارهای شناخته‌شده، استوار هستند (امضای یک بدافزار دنباله‌ای از بایت‌های درون فایل بدافزار است که منحصربه‌فرد بوده و در فایل دیگری یافت نشود). در این روش‌ها که در بین آنتی‌ویروس‌های تجاری بسیار معمول هستند، از پایگاه داده‌ای از امضاهای بدافزارها استفاده می‌کنند به این صورت که هنگام مواجهه باهر فایل، وجود یا عدم وجود امضای استخراج‌شده از آن فایل در پایگاه داده مورد بررسی قرار می‌گیرد و در صورت بروز تطابق دقیق[4] امضای فایل باهر یک از امضاهای موجود در پایگاه داده، فایل مورد نظر به‌عنوان بدافزار شناسایی می‌شود.

اگرچه روش‌های شناسایی مبتنی بر امضا دقت و کارآمدی خود را در تشخیص بدافزارهای از پیش شناسایی‌شده به اثبات رسانده‌اند، اما در شناسایی بدافزارهای جدید و ناشناخته بسیار ناتوان هستند. همچنین استفاده از روش‌های مبتنی بر امضا بسیار پرهزینه می‌باشد، چرا که استفاده از این روش‌ها مستلزم شناسایی اولیه، تحلیل دقیق، استخراج امضا و نهایتاً اضافه کردن امضا به پایگاه داده‌ها می‌باشد که انجام این مراحل دارای هزینه بسیار زیاد مالی و زمانی است و این فرصت را در اختیار بدافزار قرار می‌دهد که در سطح وسیعی گسترش‌یافته و موجب آسیب‌های فراوانی گردد.

---

[1] Signature-based
[2] Anomaly detection
[3] Heuristic-based
[4] Exact matching

۱-۲-۲- *روش‌های شناسایی ناهنجاری:*

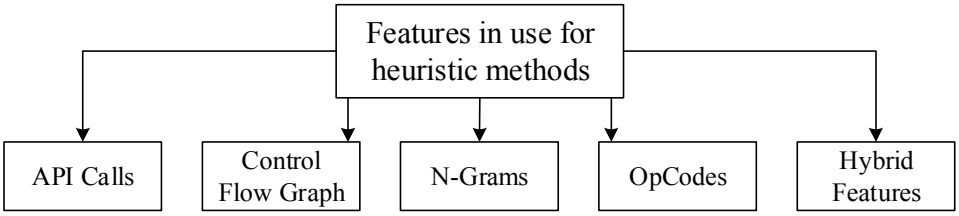
شناسایی ناهنجاری یعنی پیدا کردن الگوهای خاصی در داده‌ها که رفتار آنها با رفتاری که از آنها انتظار می‌رود مطابقت ندارد [1]. روش‌های شناسایی ناهنجاری سعی در یادگیری رفتارهای طبیعی برنامه‌های کامپیوتری دارند. پس از انجام عملیات یادگیری، در صورت مشاهده هرگونه رفتار غیرطبیعی عامل بروز آن را به‌عنوان مورد مشکوک معرفی می‌کنند. نسبت مثبت نادرست بالا را می‌توان به‌عنوان مهم‌ترین مشکل موجود در این روش‌ها نام برد.

۱-۲-۳- *روش‌های مکاشفه‌ای:*

کاستی‌های روش‌های مبتنی در امضا و ناتوانی آنها در شناسایی بدافزارهای نوین و همچنین میزان خطای بالای روش‌های شناسایی ناهنجاری باعث ظهور نسل جدیدی از تکنیک‌های شناسایی با نام روش‌های مکاشفه‌ای شد که این روش‌ها از ابزارهای یادگیری ماشین و داده‌کاوی جهت شناسایی بدافزارها استفاده می‌کنند. روش‌های مکاشفه‌ای بر پایه مجموعه‌ای وسیع از داده‌های مرتبط باهر دو کلاس (خوش‌خیم[۱] و بدافزار) استوار هستند. این روش‌ها با استفاده از ویژگی‌های ساختاری استخراج‌شده از فایل‌های خوش‌خیم و بدافزار سعی در یادگیری مدل رفتاری آنها و آموزش کلاسه کننده‌های یادگیری ماشین دارند. اگر چه فراهم کردن این حجم وسیع از داده‌های برچسب دار همیشه کار ساده‌ای نبوده و در برخی موارد غیرممکن است و همین امر به یکی از نقاط ضعف این روش‌ها تبدیل شده است، اما از مهم‌ترین مزایای این روش‌ها می‌توان به قابلیت شناسایی بدافزارهای ناشناخته و نوظهور اشاره کرد. شکل ۱ ویژگی‌های ساختاری مورد استفاده در این روش‌ها را نمایش می‌دهد.

---

[۱] Benign



شکل ۱: ویژگی‌های ساختاری مورد استفاده در روش‌های مکاشفه‌ای

این ویژگی‌های ساختاری به دو طریق، از داده‌های آموزشی استخراج می‌شوند:

۱. تحلیل ایستا

۲. تحلیل پویا

باوجود اینکه هر دو روش نحوه عملکرد بدافزار را شرح می‌دهند، اما ابزار، زمان و تخصص مورد نیاز برای هر کدام به طور چشمگیری باهم متفاوت است. تحلیل ایستا، اطلاعاتی در مورد کنترل فرایندها، دنباله‌ها و دیگر اطلاعات آماری را بدون اجرا کردن بدافزار در اختیار قرار می‌دهد. از مزایای این روش می‌توان به نداشتن سربار زمانی اجرا و عدم احتیاج به اجرا کردن بدافزار اشاره کرد. در مقابل، نیاز به دانش بسیار بالا و نادقیق بودن فرضیات در مورد نحوه عملکرد دقیق بدافزار از معایب این روش‌ها می‌باشد.

تحلیل پویا بر پایه اجرای برنامه‌ها در یک محیط محافظت‌شده، در یک ماشین واقعی و یا مجازی استوار است. در این تحلیل اطلاعات واقعی در مورد فرایند کنترل داده و دنباله‌ها در دسترس خواهد بود. اما سربار زمانی بالا و قابل‌شناسایی بودن این روش‌ها توسط بدافزارها را می‌توان به‌عنوان نقاط ضعف این روش‌ها بیان کرد.

همان طور که در شکل ۱ مشاهده می‌شود، از جمله ویژگی‌های ساختاری که در این روش‌ها مورد استفاده قرار می‌گیرند می‌توان به فراخوانی‌های توابع سیستمی[۱] ، گراف کنترل جریان[۲]، N-Gram ها، کدهای عملیاتی زبان اسمبلی[۳] و ویژگی‌های ترکیبی[۴] اشاره کرد این ویژگی‌های

---

[۱] API calls

[۲] Control flow graph

[۳] Assembly operational codes

[۴] Hybrid features

ساختاری از درون فایل‌های هر دو کلاس استخراج‌شده و به‌منظور آموزش داده کلاسه کننده‌های یادگیری ماشین مورد استفاده قرار می‌گیرند. در سال‌های اخیر و با پیشرفت تکنیک‌های مکاشفه‌ای و نیز مزایای قابل‌توجه آنها، این روش‌ها به طور گسترده مورد استفاده قرارگرفته و حتی جایگاه خود را در بین آنتی‌ویروس‌های تجاری نیز پیداکرده‌اند.

در این پایان‌نامه، یک روش نوین مکاشفه‌ای برای شناسایی بدافزارهای ناشناخته با استفاده از توزیع درون کلاسی توالی کدهای عملیاتی زبان اسمبلی که از فایل‌های اجرایی قابل‌حمل سیستم‌عامل ویندوز استخراج‌شده‌اند، ارائه شده است. این روش از Term Frequency (TF) [2] توالی کدهای عملیاتی موجود در یک فایل استفاده کرده و با معرفی یک معیار وزن دهی سعی در استخراج ویژگی‌های ساختاری مناسب برای هر دو کلاس بدافزار و خوش‌خیم دارد. سپس با استفاده از این ویژگی‌ها، کلاسه کننده‌های یادگیری ماشین را آموزش داده و با استفاده از آنها به شناسایی بدافزارهای ناشناخته می‌پردازد. دقت تشخیص بالا، سادگی و پیچیدگی زمانی کم از مهم‌ترین مزایای روش ارائه‌شده در این پایان‌نامه می‌باشد. همچنین از نسخه گسترش‌یافته این روش می‌توان برای شناسایی و دسته‌بندی خانواده بدافزارها استفاده کرد. نتایج به‌دست آمده بیانگر دقت و کارآمدی چشمگیر روش ارائه شده است.

۱-۳- ساختار رساله

ادامه این پایان‌نامه به ترتیب زیر طرح‌بندی شده است:

- **فصل دوم:** این فصل به ارائه دانش پیش زمینه و ادبیات موضوع می پردازد که مطالعه آنها، به درک بهتر مفاهیم مطرح شده در این کمک می‌کند. مطالعه این فصل برای خوانندگانی که دارای اطلاعات کافی در حوزه یادگیری ماشین هستند، ضروری نمی‌باشد.

- **فصل سوم:** در این فصل، پیشینه علمی پژوهش مورد بررسی قرار می‌گیرد. از آنجا که روش ارائه شده در این پایان‌نامه با کدهای عملیاتی زبان اسمبلی بهره می‌برد، این فصل با تمرکز بیشتر بر این مبحث تدوین شده است.
- **فصل چهارم:** در این فصل، روش انجام پژوهش به همراه جزئیات کامل مورد بررسی قرار می‌گیرد. در این فصل سعی شده است که با ذکر مثال و ارائه اشکال گویا، به شرح الگوریتم پیشنهادی پرداخته شود.
- **فصل پنجم:** در این فصل مجموعه داده‌هایی که برای ارزیابی الگوریتم پیشنهادی مورد استفاده قرار گرفته‌اند، معرفی شده و آزمایش‌ها و نتایج انجام شده بر روی الگوریتم پیشنهادی نمایش داده‌شده‌است.
- **فصل ششم:** در این فصل جمع بندی مطالب، نتیجه گیری، و پیشنهاداتی برای آینده ذکر شده است.