پنجمین کنفرانس ملی مباحث نوین در کامپیوتر و فناوری اطلاعات
5th National Conference on Advanced Topics in Computer and Information Technology
۱۴ آذر ماه ۱۴۰۰

استفاده از الگوریتم‌های فراابتکاری در شناسایی بدافزار

عاطفه محمدی

Mohammadi289@yahoo.com

خلاصه: امروزه افزایش روز به روز بدافزارها و به‌دنبال آن یافتن راه‌های مناسب، یکی از مهمترین چالش‌های امنیت اطلاعات و شبکه‌های ارتباطی جهت حفاظت سیستم‌ها در برابر بدافزارها بوده که شناخت به موقع بدافزارها و یافتن راه‌های مقابله با اثرات مخرب آنها از مهمترین دغدغه‌های برنامه‌نویسان و متخصصین امنیت اطلاعات می‌باشد. در سال‌های اخیر محققان تلاش‌های بسیاری در جهت تشخیص بدافزار انجام داده‌اند که در این راستا نیز تحقیقاتی در زمینه استفاده از تکنیک‌های داده کاوی و هوش مصنوعی صورت گرفته است.در این مقاله سعی بر این است تا با استفاده از الگوریتم‌های داده کاوی و فراابتکاری به تشخیص فایل‌های آلوده به بدافزار بپردازیم.

کلمات کلیدی: بدافزار، داده کاوی، الگوریتم‌های فرابتکاری، هوش مصنوعی.

۱ – مقدمه

در سال‌های اخیر، اینترنت به عنوان بخشی از زندگی روزانه مردم تبدیل شده است چرا که خدمات متنوع، توسط این شبکه ارتباطی به کاربران عرضه می‌گردد. اینترنت از یک شبکه ارتباطی به شبکه‌ای از منابع ارائه دهنده اطلاعات، شکل تازه‌ای از تعاملات اجتماعی و بازاری برای فروش محصولات و ارائه سرویس‌ها و خدمات متنوع از جمله بانکداری اینترنتی و تبلیغات تبدیل گردیده است. در واقع رشد فزاینده و استفاده همگانی افراد از اینترنت به عنوان یک سوژه جذاب برای انجام کارهای خرابکارانه توسط مهاجمان تبدیل شده و انگیزه آنان برای نفوذ در شبکه و ایجاد حملات اینترنتی گسترده‌تر، همواره در حال شدت یافتن است. بدافزار مخرب، یکی از مهمترین و اضطراری‌ترین تهدیدات امنیتی پیش روی اینترنت است و شرکت‌های ضدویروس معمولا باید هر روز برای مقابله با ده‌ها هزار نمونه بدافزار جدید آماده باشند[۱،۲].

رشد سریع بدافزار سبب بوجود آمدن تهدیدهای بسیار زیادی در حوزه امنیت اطلاعات شده است. بنابراین مراکز دفاع سایبری که همانند مرزهای یک کشور هستند از اهمیت زیادی برخوردار بوده و در کنار فضای مجازی می‌توانند مورد تهدید و هجوم قرار بگیرند. نرم‌افزارهای مخرب در طول زمان از یک مزاحمت غیرجدی تبدیل به یک تهدید امنیتی جدی شده‌اند که این امر سبب شده است توجه محققان امنیتی در سراسر جهان به این موضوع جلب شود. بنابراین به منظور تعیین رفتار، حالت وخیم بودن و عملیات انواع نرم‌افزارهای مخرب؛ باید مورد تجزیه و تحلیل قرار بگیرند [۳].

۲ – کارهای پیشین

در طول سال‌های گذشته، روش‌های متنوعی به منظور شناسایی فعالیت‌های بدافزارها انجام شده است که در ادامه به اختصار به چند مورد اشاره می‌شود:

جدول ۱: کارهای پیشین

روش پیشنهادی	مرجع	دقت
استفاده از طبقه‌بندهای NaiveBayes، Logistic Regression و SVM	[۴]	۹۵،۵۳
Naive Bayes	[۵]	۷۹،۹۷
ماشین بردارپشتیبان الگوریتم ازدحام ذرات	[۶]	۹۰
ماشین بردار پشتیبان	[۷]	۹۹،۷۰
ماشین بردار پشتیبان	[۸]	۸۵
ماشین بردار پشتیبان مبتنی بر تابع شعاعی	[۹]	۹۲

۳ –داده‌کاوی

داده‌کاوی به فرآیند استخراج خودکار مدل‌ها از میان انبوه داده‌ها اطلاق می‌شود. اخیراً پیشرفت‌های قابل توجهی در زمینه داده‌کاوی به دست آمده است و داده‌کاوی مجموعه الگوریتم‌های گسترده‌ای را از علومی چون آمار، تشخیص الگو، یادگیری ماشین و پایگاه داده‌ها بکار گرفته است. از رایج‌ترین مدل‌های داده‌کاوی طبقه‌بندی می‌باشد، که می‌تواند با استفاده از نمونه‌های از قبل طبقه‌بندی شده برای توسعه یک مدل، تعداد زیادی نمونه را طبقه‌بندی کند. فرایند طبقه‌بندی داده‌ها شامل یادگیری و طبقه‌بندی می‌باشد. در یادگیری، داده‌های آموزشی توسط الگوریتم طبقه‌بندی تجزیه و تحلیل می‌شوند اما در طبقه‌بندی، داده‌های ارزیابی برای تخمین دقت طبقه‌بندی مورد استفاده قرار می‌گیرند. اگر دقت تخمین زده شده توسط طبقه‌بندی قابل قبول باشد، می‌توان مجموعه داده‌های جدید را نیز به طبقه‌بندی اعمال کرد[۱۱،۱۰]. برای بکارگیری تکنیک‌های داده کاوی در سیستم‌های تشخیص بدافزار، متخصصان داده کاوی روش‌های مختلفی را بکار می‌گیرند[۱۳،۱۲].

۴– الگوریتم‌های فراابتکاری

الگوریتم‌های فراابتکاری یکی از انواع الگوریتم‌های ابتکاری هستند که در مسائل گوناگون کاربرد دارند. این الگوریتم‌ها که نوعی از الگوریتم‌های تصادفی هستند برای یافتن پاسخ بهینه به کار برده می‌شوند. بخش عظیمی از این الگوریتم‌ها که از زندگی اجتماعی حشرات الهام می‌گیرد بسیار قدرتمند و قوی هستند و در برخی کلونی آنها بطور مداوم در حال رشد و توسعه می‌باشند و یا در برخی از آنها بر اثر مسائلی مانند شکار کاهش می‌یابند و یا بر اثر تقسیم کلونی تجزیه می‌شوند. اما چیزی که مهم است این است که بطور معمول حشرات با تغییر شرایط در عرضه مواد غذایی و عادت‌های خود کنار می‌آیند. این حشرات بسیار کوچک بوده و به تنهایی قدرت آنچنانی ندارند اما جمع آنها بسیار قوی و قدرتمند خواهد بود. در ادامه به شرح مختصری از برخی الگوریتم‌های فراابتکاری خواهیم پرداخت.

- **الگوریتم‌سنجاقک**(DA)

ایده اصلی الگوریتم سنجاقک از رفتار گروهی سنجاقک‌ها در حالت ایستا (استراحت) و پویا (حرکت یا جنب‌وجوش) در طبیعت الهام گرفته است. دو مرحله اساسی در بهینه‌سازی، اکتشاف و بهره‌برداری است که از کار گروهی در راهبری، جستجو غذا و اجتناب از دشمنان در حالت‌های ایستا و پویا الگوبرداری شده است.

- **الگوریتم‌بهینه‌سازی‌نهنگ‌یا وال**(WOA)

ساختار این الگوریتم از شیوه شکار حباب خالص وال‌ها الهام گرفته است. نهنگ کوهان‌دار ترجیح می‌دهد تا گروهی از کریل‌ها یا ماهیان کوچک که نزدیک به سطح آب هستند را شکار کند. مدل ریاضی الگوریتم بهینه‌سازی نهنگ که مبتنی بر روش تغذیه حباب خالص است شامل مراحل محاصره طعمه، مانور تغذیه حباب خاص به صورت مارپیچی و به دام انداختن شکار می‌باشد.

- **الگوریتم‌بهینه‌سازی‌گرگ‌خاکستری**(GWO)

الگوریتم گرگ خاکستری یک الگوریتم فراکاوشی الهام گرفته از طبیعت است که اساس آن بر پایه ساختار سلسله مراتبی و رفتار اجتماعی گرگ‌ها در زمان شکار می‌باشد. این الگوریتم فرایند ساده‌ای را در تنظیمات دارد و به راحتی قابلیت تعمیم به مسایل با ابعاد بزرگ را دارا می‌باشد. در پیاده‌سازی این الگوریتم، چهار نوع از گرگ‌های خاکستری مانند آلفا، بتا، دلتا و امگا برای شبیه‌سازی سلسله مراتب رهبری استفاده شده است که در آن سه گام اصلی از شکار، جستجو برای طعمه، محاصره طعمه و حمله به طعمه، اجرا می‌شوند. این الگوریتم تنها دارای دو پارامتر تنظیم می‌باشد.

- **الگوریتم بهینه‌سازی ملخ**(GOA)

این الگوریتم با الهام از رفتار اجتماعی ملخ‌ها و نحوه تأثیرپذیری هر ملخ از محیط پیرامونش طراحی شده است. در این الگوریتم بروزرسانی موقعیت هر ملخ به فاصله هر ملخ از تمام جمعیت ملخ‌ها در نسل جاری و موقعیت بهترین ملخ وابسته است. از ویژگی‌های این الگوریتم می‌توان به سادگی و دارا بودن فقط یک پارامتر تنظیم، ارجاع کرد.

- **الگوریتم جستجوی کلاغ‌ها**(CSA)

این الگوریتم از رفتار هوشمند کلاغ‌ها الهام گرفته و همچنین یک تکنیک مبتنی بر جمعیت است. ایده اصلی کلاغ این است که کلاغ‌ها غذای مازاد خود را مخفی کرده و ذخیره می‌کنند و در صورت نیاز آن را پیدا می‌کنند. کلاغ‌ها می‌توانند از ابزارها استفاده کنند و با روش‌های پیچیده‌ای ارتباط برقرار کنند و مکان‌های مخفی غذا خود را تا چند ماه بعد به یاد آورند. آنها مکان‌هایی را که پرندگان دیگر غذای خود را پنهان می‌کنند، مشاهده می‌کنند و زمانی که صاحب آن محل را ترک می‌کند آنها را سرقت می‌کنند.

۵ –شرح روش پیشنهادی

در روش پیشنهادی پس از پیش‌پردازش بر روی مجموعه داده موردنظر گردآوری شده از سایت kaggle[1]، جهت طبقه‌بندی فایل‌ها به دو مجموعه فایل‌های آلوده به بدافزار و فایل‌های سالم از الگوریتم‌های داده کاوی استفاده می‌شود.

باتوجه به مجموعه داده موردنظر در این تحقیق پیش‌پردازش شامل نرمال‌سازی داده‌ها می‌باشد. برای این منظور در این مقاله برای نرمالیزه کردن داده‌ها از روش نرمالیزاسیون آماری $\max - \min$، در بازه $[1,0]$ که از رابطه ۱ تبعیت می‌کند استفاده کردیم :

$$X_{norm} = \frac{x-\min(x)}{\max(x)-\min(x)} \qquad (۱)$$

که x مقدار داده مورد نظر جهت نرمال شدن، $\min(x)$ کمینه بردار ورودی x، $\max(x)$ بیشینه بردار ورودی x و X_{norm} مقدار نرمال شده x می‌باشد.

در گام بعدی جهت ارزیابی داده‌ها با استفاده از روش هولداوت داده‌ها به دو بخش آموزشی و ارزیابی تقسیم می‌شود(۷۰ درصد مجموعه داده برای آموزش و ۳۰ درصد آن برای آزمایش مورد استفاده قرار می‌گیرد).

پس از تقسیم‌بندی مجموعه داده، در گام بعدی شبکه عصبی مصنوعی از نوع پیشخور ایجاد می‌شود. پس از ایجاد شبکه عصبی پیشنهادی و تقسیم‌بندی داده‌ها، شبکه عصبی آموزش داده شد. از مهمترین اهداف شبکه عصبی مصنوعی یافتن وزن‌های متناسب با نرون‌ها در لایه‌های مختلف است. فرایند تعیین وزن‌ها بدین صورت است که در ابتدا وزن‌های شبکه عصبی بصورت تصادفی مقداردهی شده و در هر بار تکرار تغییر پیدا می‌کنند و تا زمانی که اختلاف خطای پیش‌بینی شبکه عصبی و خروجی واقعی متناظر با ورودی‌ها، کمتر از یک حد معین شود، اصلاح می‌شوند. در این تحقیق جهت تعیین وزن‌های شبکه عصبی پیشنهادی از الگوریتم‌های نهنگ، سنجاقک، گرگ خاکستری، کلاغ، مورچه و ملخ استفاده شد.

۶– ارزیابی روش پیشنهادی

در این تحقیق برای سنجش و ارزیابی عملکرد روش پیشنهادی براساس ماتریس درهم‌ریختگی از معیارهای دقت طبق رابطه ۲ استفاده می‌شود:

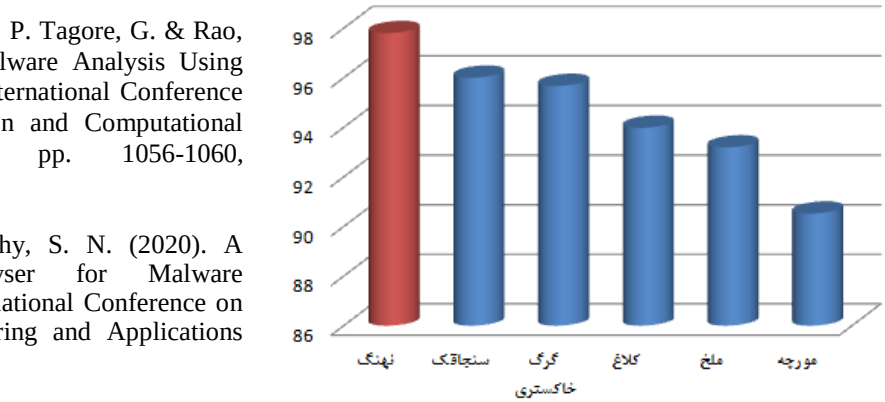
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \qquad (۲)$$

دقت بدست آمده با استفاده از الگوریتم‌های مورد استفاده برای داده‌های آموزشی براساس درصد در جدول ۲ بیان گردیده است.

جدول ۲: نتایج بدست آمده

مورچه	ملخ	کلاغ	گرگ خاکستری	سنجاقک	نهنگ
90.52	**93.19**	**93.98**	**95.67**	**96**	**97.8**

[1] www.kaggle.com/nsaravana/malware-detection



شکل ۱: نتایج مورد مقایسه

باتوجه به نتایج بدست آمده مشاهده می‌کنیم که الگوریتم نهنگ نسبت به سایر الگوریتم‌های مورد بررسی از دقت بالاتری برخوردار است.

۷- نتیجه گیری

در این تحقیق روشی جهت تشخیص بدافزارها با استفاده از الگوریتم‌های شبکه عصبی مصنوعی و الگوریتم‌های فراابتکاری ارائه شد. با توجه به پیاده‌سازی روش پیشنهادی و نتایج بدست آمده می‌توان گفت الگوریتم‌های فراابتکاری با بهبود شبکه عصبی مصنوعی از عملکرد خوبی جهت تشخیص فایل‌های آلوده به بدافزار برخوردار هستند.

مراجع

[1] Laurenza., Giuseppe., et al, "An Architecture for Semi-Automatic Collaborative Malware Analysis for CIs." Dependable Systems and Networks Workshop, 46th Annual IEEE/IFIP International Conference on. pp. 137-142. IEEE 2016.

[2] Hardy., William., et al. "DL4MD.,A Deep Learning Framework for Intelligent Malware Detection.",Proceedings of the International Conference on Data Mining (DMIN). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2016.

[3] Jamalpur, S. Navya, Y.S. Raja, P. Tagore, G. & Rao, G.R. (2018), «Dynamic Malware Analysis Using Cuckoo Sandbox, Second International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 1056-1060, Coimbatore, India.

[4] Samantray, O. P., & Tripathy, S. N. (2020). A Knowledge-Domain Analyser for Malware Classification. In 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA) (pp. 1-7). IEEE.

[5] Anggraini, I., Kunang, Y. N., & Firdaus, F. (2020). Penerapan Naive Bayes Pada Detection Malware dengan Diskritisasi Variabel. Telematika, 13(1), pp.11-21

[6] آرایش، م،۱۳۹۷،کشف بدافزار اسب تروا با استفاده از تکنیک‌های داده کاوی ،ص۷۷

[7] Cuan, B, Damien, A, Delaplace, C, Valois, M, 2018, Malware Detection، in PDF Files Using Machine Learning, In Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (ICETE 2018) - Volume 2: SECRYPT, pages 412-419

[8] Rehman, Z. U., Khan, S. N., Muhammad, K., Lee, J. W., Lv, Z., Baik, S. W., & Mehmood, I, "Machine learning-assisted signature and heuristic-based detection of malwares in Android devices", Computers & Electrical Engineering, vol.69, pp.828-841. 2018.

[9] Jain, G., Raghuwanshi, S., & Vishwakarma, G. "Hardware Trojan: Malware Detection Using Reverse Engineering and SVM". In International Conference on Intelligent Systems Design and Applications, pp. 530-539. Springer, Cham. 2017.

[10] Chi-Chen, An-An Chiu, Shaio Yan Huang, David C. Yen, "Detecting the financial statement fraud: The analysis of differences between data mining techniques and experts judgments", Knowledge-Based Systems, Vol. 89, (2015), pp. 459–470.

[11] V.P. Darabad, M. Vaklilan, T.R. Blackburn, B.T.

Phung, "An efficient PD data mining method for power transformer defect models using SOM technique", International Journal of Electrical Power & Energy Systems, Vol. 71, (2015), pp. 373–382.

[12] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). "Data Mining: Practical Machine Learning Tools And Techniques". Morgan Kaufmann.

[13] Ye, Y., Li, T., Adjeroh, D. And Iyengar, S.S., (2017). "A Survey On Malware Detection Using Data Mining Techniques". ACM Computing Surveys (CSUR), 50(3), P.41.