PROVIDING CYBER SECURITY SOLUTION FOR MALWARE DETECTION USING SUPPORT VECTOR MACHINE ALGORITHM (SVM)

Narmada B, Arfath Khan M, Mahalakshmi P, Jayasree S

Narmada B, Assistant Professor and Head of the Department, Department of Computer Science and Engineering, Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India
Arfath Khan M, Student, Department of Computer Science and Engineering, Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India
Mahalakshmi P, Student, Department of Computer Science and Engineering, Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India
Jayasree S, Student, Department of Computer Science and Engineering, Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India

---***---

Abstract – *Malware detection developers faced an issue with a generation of recent signatures of malware code. A very famous and recognized technique is the pattern-based malware code detection technique. This results in the evasion of signatures that are built to support the code syntax. During this paper, we discuss some well-known methods of malware detection supported by the semantic feature extraction technique. In the current decade, most of the authors focused on the malware feature extraction process for the generic detection process. The effectiveness of the Malicious Sequence Pattern Matching technique for malware detection invites moderation and improvement of the present system and method. Some authors used the rule mining technique, another used the graph technique and a few also focused on the feature clustering process of malware detection. The focus of the Multi-Classification framework is to detect the malicious affected files. To protect legitimate users from attacks, the foremost significant line of defense against malware is anti-malware software products, which mainly use signature-based methods for detection. Machine Learning algorithms are proved useful at identifying zero-day attacks or detecting an unusual behavior of systems that might indicate an attack or malware.*

Key Words: **Data Processing, Computer Science, Cyber Security, J48, SVM Algorithm, KDD, Malware Detection**

1. INTRODUCTION

1.1 DATA MINING

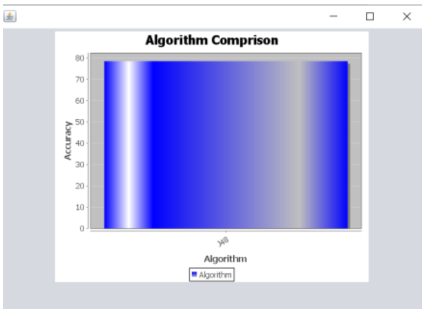
Data mining is an interdisciplinary subfield of engineering science. It's the computational process of discovering patterns in large data sets involving methods at the intersection of computing, machine learning, statistics, and database systems. The goal of the information mining process is to extract information from an information set and transform it into a lucid structure for further use. Except for the raw analysis step, it involves database and data management aspects, data processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data processing is the analysis step of the "knowledge discovery in databases" process or KDD.

1.2 Cyber Security

The cyberattack surface in modern enterprise environments is huge, and it is continuing to grow rapidly. This implies that analyzing and improving an organization's cybersecurity posture needs over mere human intervention. AI is now becoming essential to information security, as these technologies are capable of swiftly analyzing numerous data sets and tracking down a large sort of cyber threats. These technologies are continually learning and improving, drawing data from past experiences and the present to pinpoint new sorts of attacks that will occur today or tomorrow.

2. EXISTING SYSTEM

Due to its damage to Internet security, malware(e.g., virus, worm, Trojan)and its finding has caught the eye of both the anti-malware industry and researchers for many years. To shield genuine users from the attacks, the foremost significant line of defense against malware is anti-malware software products, which mostly use signature-based methods for detection. However, this method fails to acknowledge new, unseen malicious executables. To unravel this problem, during this paper, supported the instruction sequences extracted from the file sample set, we propose a good sequence mining algorithm to get malicious sequential patterns, then J48 classifiers constructed for malware detection supported the discovered patterns. The developed data processing framework composed of the proposed sequential pattern mining method and J48 classifier can well characterize the malicious patterns from the collected file samples to effectively detect newly unseen malware samples.



2.1 Drawbacks

✓ There isn’t any mechanism to predict the unseen malware file.

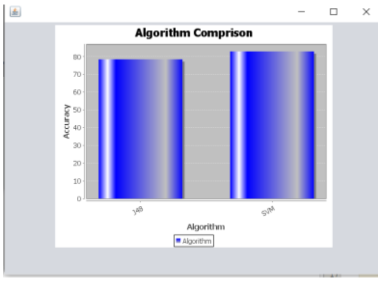
✓ Traditional signature-based anti-virus systems fail to detect unseen malicious executables.

✓ The J48 method learns over sequences of hard and fast length.

✓ It doesn’t find the sort of malware file.

3. PROPOSED SYSTEM

Sequence mining algorithm to get malicious sequential patterns supported the machine instruction sequences extracted from the Windows Portable Executable (PE) files, then use to construct a knowledge mining framework, called MSPMD (**M**alicious **S**equential **Pattern-based** **M**alware **D**etection), to detect new malware samples. The most contributions of this paper are summarized as follows: Instruction sequences are extracted from the PE (Portable Executable) files because the preliminary features, that supported the malicious sequential patterns are mined in the next step. The extracted instruction sequences can well indicate the potential malicious patterns at the microlevel. Additionally, such reasonable features will be easily extracted and accustomed to generating signatures for the standard malware detection systems. A good sequential pattern mining algorithm, called MSPE (Malicious **S**equential **P**attern **E**xtraction), to find malicious sequential patterns from the instruction sequence. MSPE introduces the concept of objective-oriented to find out patterns with strong abilities to differentiate malware from benign files. Moreover, we design a filtering criterion in MSPE to filter the redundant patterns within the mining process and to scale back the prices of processing time and search space. This strategy greatly enhances the efficiency of our algorithm. SVM classifier for malware detection: We propose an SVM classifier as a detection module to spot malware. Different from the normal *k*-nearest-neighbor method, SVM chooses *k* automatically during the algorithm process. More importantly, the SVM classifier is well-matched with the discovered sequential patterns and is in a position to get better results than other classifiers in malware detection. To conduct a series of experiments to judge each part of our framework and therefore the whole system supported real sample collection, containing both malicious and benign PE files. The results show that MSPMD is a good and efficient solution for detecting new malware samples.



3.1 Advantages

✓ A Classification is predicated on the generated rules.

✓ It is ready to automatically generate strong signatures.

✓ IItssignatures are very useful for malware detection.

✓ To simulate the task of detecting new malicious executables.

✓ It is in a position to extract behavioral features from a novel structure in portable executable

4. SYSTEM REQUIREMENTS

4.1 Hardware Requirements

Processor: Any Processor above 500 MHz

RAM: 128MB.

Hard Disk: 10 GB.

Compact Disk: 650 MB.

Input device: Standard Keyboard, Mouse.

Output device: VGA Monitor

4.2 Software Requirements

Operating System: Windows OS

Front End: JAVA

5. SYSTEM IMPLEMENTATION

5.1 Module Split up

✓ Data samples Acquisition
✓ Instruction Sequence Extractor
✓ Malicious Sequential Pattern Miner
✓ NNaiveBayes Classifier

5.2 Modules Description

5.2.1 Data samples acquisition

This module is employed to input to the system. It contains malicious and benign files. Within the training phase, a classifier is generated malicious sequence pattern mining based on malicious and benign.

5.2.2 Instruction Sequence Extractor

Instruction sequences are extracted from the PE (Portable Executable) files because the preliminary features, support which the malicious sequential patterns are mined within the next step. The extracted instruction sequences can well indicate the potential malicious patterns at the microlevel. Additionally, such quiet features will be easily extracted and used to generate signatures for the normal malware detection systems

5.2.3 Malicious Sequential Pattern Miner

An effective sequential pattern mining algorithm, called MSPE (Malicious **S**equential **P**attern **E**xtraction), to get malicious sequential patterns from the instruction sequence. MSPE introduces the concept of objective-oriented is told patterns with strong abilities to differentiate malware from benign files. Moreover, we design a filtering criterion in MSPE to filter the redundant patterns in the mining process and to scale back the prices of processing time and search space.

5.2.4 Naive Bayes Classifier

Naive Bayes classifier for malware detection: we propose a naive Bayes classifier as a detection module to spot malware. The naive Bayes classifier is well-matched with the discovered sequential patterns and is ready to get better results than other classifiers in malware detection. The results show that MSPMD is an effective and efficient solution for detecting new malware samples.

5.2.5 Evaluation criteria

In this module, the performance of the proposed Support Vector Machine(SVM) algorithm is extensively compared with thereupon of some existing J48 (Part of Decision Tree)algorithms. To research the performance of various algorithms, the experimentation is completed on Sequence data patterns of Software applications. The most important metrics for evaluating the performance of various algorithms are the category separability index and detection accuracy of the support vector machine rule. The proposed system provides an improved accuracy rate in malware detection.

6. CONCLUSION

To develop a data-mining-based detection framework called Malicious Sequential Pattern-based Malware Detection (MSPMD), which consists of the proposed sequential pattern mining algorithm (MSPE) and SVM classifier. It first extracts instruction sequences from the PE file samples and conducts feature selection before mining; then MSPE is applied to come up with malicious sequential patterns. For the testing file samples, after feature representation, an SVM classifier is made for malware detection. Unlike the previous research which is unable to mine discriminative features, we propose to use a sequence mining algorithm on instruction sequences to extract well representative features.

.6.1 Future work

Future work includes partitioning the first gene set into some distinct subsets or clusters so that the malware within a cluster is tightly coupled with a strong association to the sample categories. We will extend the work to implement various classification algorithms to enhance the accuracy rate at the time of malware prediction.

7. REFERENCES

- [1] 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) | 978-1-7281-7089-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICISS49785.2020.931600

- [2] Narouei,M. , Ahmadi, M., Giacinto, G., Takabi, H.,&Sami, A. (2015). DLL Miner: Structural mining for malware detection. Security and Communication Networks, 8.

- [3] Ye, Y., Wang, D., Li, T., Ye, D., & Jiang, Q . (2008) . An intelligent PE-malware detection system based on association mining. Journal in computer virology, 4, 323–334.

- [4] Ye, Y., Li, T., Chen, Y., & Jiang, Q. (2010). Automatic malware categorization using cluster ensemble. In Proceedings of the 16th international conference on knowledge discovery and data mining (pp.95–104).

- [5] Wchner, T., Ochoa, M. , & Pretschner, A. (2014). Malware detection with quantitative data flow graphs. In Proceedings of the 9th ACM symposium on information, computer and communications security (pp.271–282).

- [6] Nissim, N. , Moskovitch, R. , Rokach, L. & Elovici, Y. (2014). A Noval approach for active learning methods for enhanced PC malware detection in windows OS. Expert Systems with Applications, 41,5843–5857 •

- [7] P. Prajapati and P. Shah, “A review on secure data deduplication: Cloud storage security issue,” Journal of King Saud University-Computer and Information Sciences, 2020.

- [8] "Cyber security breaches report of black hat ethical hacking," 2019, [online] https://www.blackhatethicalhacking.com
- [9] "Verizon data breach investigations report," 2020, [online] https://enterprise.verizon.com/resources/reports/2020- data-breachinvestigations-report.pdf
- [10] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, and V. C. Leung, "A survey on security threats and defensive techniques of machine learning: A data-driven view," IEEE Access, vol. 6, pp. 12 103–12 117, 2018.
- [11] P. Prajapati, P. Shah, A. Ganatra, and S. Patel, "Efficient cross user client-side data deduplication in Hadoop." JCP, vol. 12, no. 4, pp. 362–370, 2017.
- [12] P. Prajapati and P. Shah, "Efficient cross user data deduplication in remote data storage," in International Conference for Convergence for Technology-2014. IEEE, 2014, pp. 1–5.
- [13] S. Kadvani, A. Patel, M. Tilala, P. Prajapati, and P. Shah, "Provable data possession using identity-based encryption," in Information and Communication Technology for Intelligent Systems. Springer, 2019, pp. 87–94.
- [14] P. Shah and P. Prajapati, "Provable data possession using additively homomorphic encryption," Journal of King Saud University-Computer and Information Sciences, 2020.
- [15] P. Prajapati, K. Dave, and P. Shah, "A review of recent blockchain applications," International Journal of Scientific & Technology Research, vol. 9, pp. 897–903, 2020.
- [16] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, and C. Wang, "Machine learning and deep learning methods for cybersecurity," IEEE Access, vol. 6, pp. 35 365–35 381, 2018.
- [17] P. Wei, Y. Li, Z. Zhang, T. Hu, Z. Li, and D. Liu, "An optimization method for intrusion detection classification model based on deep belief network," IEEE Access, vol. 7, pp. 87 593–87 605, 2019.
- [18] Y. Li, K. Xiong, T. Chin, and C. Hu, "A machine learning framework for domain generation algorithm based malware detection," IEEE Access, vol. 7, pp. 32 765–32 782, 2019.
- [19] R. K. Malaiya, D. Kwon, J. Kim, S. C. Suh, H. Kim, and I. Kim, "An empirical evaluation of deep learning for network anomaly detection," in 2018 International Conference on Computing, Networking, and Communications (ICNC). IEEE, 2018, pp. 893– 898.
- [20] F. Liang, W. G. Hatcher, W. Liao, W. Gao, and W. Yu, "Machine learning for security and the internet of things: the good, the bad, and the ugly," IEEE Access, vol. 7, pp. 158 126–158 147, 2019.