e-ISJN: A4372-3114
p-ISJN: A4372-3115

ISSN: 2321-7782 (Online)
ISSN: 2347-1778 (Print)
Impact Factor: 7.327

Volume 8, Issue 2, February 2020

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

A Survey of Machine Learning Techniques for Identifying and Classifying Malwares

Umesh V. Nikam[1]
Department of Computer Science & Engineering,
P. R. M. I. T&R, Badnera.
Amravati, India

Dr. V. M. Deshmukh[2]
Department of Computer Science & Engineering
P. R. M. I. T&R, Badnera.
Amravati, India

Abstract: A serious threat on the internet today is a malware. As the malware propagate they change their code. Nowdays attacker creates polymorphic and metamorphic malwares. The traditional signature based detection techniques are inefficient against modern day's malware threats. The various malware families have different behavior pattern reflecting their origin and purposes. These patterns can be used to detect and classify unknown malwares into their families using machine learning technique. This survey paper provides an overview of various techniques for detecting and classifying malwares into their respective families.

Keywords: Malware, Machine learning, Classification.

I. INTRODUCTION

A malware is a computer program with the purpose of causing harm to the operating system. Basic purpose of malware is to fulfill the harmful intent of an attacker by gathering personal information about a user or host system, thus hampering availability, integrity and privacy of user's data. There is a wide a range of malwares like Worm, Virus, Trojan horse, Rootkit, Backdoor, Botnet, Spyware, Adware etc.

Known software threats can be detected by modern antivirus software effectively but is inefficient in detecting novel malware. A study by AusCERT found that 80 percent of new malware was not detected by latest antivirus software. [1] Detection, mitigation and classification of malware is a major problem in internet today. The malwares are continuously growing in volume, variety and velocity.

A. LIMITATIONS OF TRADITIONAL ANTIVIRUS

Traditional signature based antivirus system is reactive in nature. In order to detect a malware in earlier days malware analyst used to manually generate a signature or a hash, and creates a database of a those signatures. During every new scan antivirus system scans the database and if there is a match detects the malware. But because of polymorphic nature of malwares; this signature based detection technique is not able to identify various security threats. In order to create a more reliable and robust system we need to develop an alternative to the traditional signature based detection system.

To overcome the drawback of signature based system, malware analysis techniques are being followed, which can be either static or dynamic. These malware analysis techniques help the analyst to understand risk associated with malicious code.

In static analysis malicious software's are analyzed without being executed. Before doing static analysis it is necessary to unpack and decrypt executables. The detection pattern used can be Byte Sequence, N Grams, Syntactic Library Call, Control Flow Graph, String Signature etc.

In Dynamic analysis malicious software are analyzed while it is being executed in a controlled environment. Before the execution of malware samples, monitoring tools like Wireshark, Regshot etc. are installed and activated.

II. Machine Learning for Detecting and Classifying Malwares

Various machines learning techniques like Association rule, Support Vector machine, Decision Tree, Random Forest, Naive Bayes and clustering have been proposed for detecting and classifying unknown samples into either known families or to identify the unseen behavior for detailed analysis. A few publications are discussed below:

Natraj et al. [2] proposed a method for classifying malwares using image processing technique. A K nearest neighbor algorithm with Euclidean distance method is used for malware classification. It is found that this technique is fast as compared to other technique but an attacker can develop an alternative technique to beat the system.

Kong et al. [3] proposed an automated framework for malware classification based on function call graph. By applying a discriminate distance metric learning technique malware samples belonging to the same family were clustered and others were separated by a marginal distance.

Tian et al. [4] proposed a model for classifying malwares using function length frequency to classify Trojans. Their result shows that function length along with its frequency is very effective in identifying malware family and if combined with other features can give fast results.

Siddiqui et al. [5] used decision tree and random forest machine learning model for malware classification. For detecting a worm in the wild they used variable length instruction sequence. A dataset of 2774 containing 1444 worms and 1330 benign files were used to test their methods.

Rieck et al. [6] proposed a framework which uses sandbox environment to monitor a malware behavior using machine learning algorithms. They used clustering to identify novel classes of malware with similar behavior and classification for unknown malwares.

Bayer et al. [7] proposed a technique based on Anubis [8] which generates execution traces of all the samples. Taint propagation capability is used along with Anubis.

Tian et al. [9] extracted API call sequences from executables using automated tool. They used classifiers from WEKA library to distinguish clean files and malware files. A dataset of 1368 malwares and 456 cleanwares was used to demonstrate their work with an accuracy of 97%.

Biley et al. [10] pointed out that malwares characterize by antivirus product are not consistent across various antivirus products and are not concise in their semantics. They created a technique for classification that describes malware behavior in terms of system state changes. Binaries are executed in a virtualized environment where a virtual machine is partially firewalled to limit impact of any immediate normalized compression distance (NCD) as a distance metric is used to cluster the malware. This technique was applied for classification and analysis of 3700 malware samples. They also compared consistency, completeness and conciseness of the cluster with AV products.

Santos [11] et al. proposed OPEM, a hybrid malware detector. This uses features from static and dynamic analysis of malicious code. By considering some learning algorithms like Decision Tree, K nearest neighbor, Bayesian network and SVM he validated the model and found that the performance is enhanced.

Islam et al. [12] used same technique of OPEM. He used function length frequency and printable string information as a static feature and API function name and API parameter as a dynamic feature. The model was tested using 2939 executable files including 541 clean files separately for every feature. The result shows that metaclassifiers achieves highest accuracy for integrated features and meta RF is the best performer of all.

Anderson [13] proposed a method which uses multiple data sources. He used multiple kernel learning technique to find weighted combination of data sources. A SVM classifier was used to classify dataset into malicious and benign. Proposed method was tested on a dataset of 780 malwares and 776 benign instances with an accuracy of 98.07%.

Table I. Various classification techniques.

S. N	Author	Algorithm Used	Classification Technique	Features	Technology
1	Natraj et. al [2]	K nearest neighbour	Eucledean distance	Fast but attacker can adopt countermeasures to beat a system.	Image processing
2	Kong et. al [3]	Distance metric learning	Function call graph	Can very well classify malwares into respective family	Data mining
3	Tian et. al [4]	Matching learning algorithm in WEKA	Function length frequency	Significant in identifying malware family & can be combined with other features for fast classification	Machine Learning
4	Siddiqui et al [5]	Decision tree, Random Forest	Variable length instruction sequence	Achieved accuracy of 92% in detecting worms.	Machine Learning
5	Rieck et. al [6]	A framework using sandbox environment	Clustering & Classification	Clustering is used to identify novel classes of malware	Machine Learning
6	Bayer et. al [7]	Clustering algorithm based on LSH	Anubis with taint propagation capability	A set of 75000 malwares were clustered in 3 hrs.	Machine Learning
7	Tian et. al [9]	Classifier from WEKA	API call sequence in virtual environment	Achieved accuracy of 97%	Machine Learning
8	Biley et. al [10]	Developed a framework for classification	Normalized compression distance(NCD)	Environment of virtualized system is static throught the experiment.	Machine Learning
9	Santos et. al [11]	Decision Tree K nearest neighbor Bayesian network,SVM	Static features & Dynamic features	Hybrid approach enhances performance when run seperately	Machine Learning
10	Islam et. al [12]	Integrated method for SVM, DT&Rf	Static features & Dynamic features	Classifiers achieve higher accuracy for integrated features & meta RF is best performer	Machine Learning
11	Anderson et. al [13]	SVM classifier	Multiple kernel learning	Accuracy achieved is 98.07	Machine Learning

III. Conclusion

Malwares are posing severe threat to internet today. Conventional antivirus products can detect only those malwares which are previously identified and their signature is available in a database. To detect and counter obfuscated malware, it is very necessary to research new techniques incorporating combination of static and dynamic features for detecting and classifying malwares. This paper highlights the existing techniques for analyzing, detecting and classifying malwares.

References

1. Kotadia. Eighty Percent of New Malware Defeats Antivirus, July 2006
2. Natraj, L. , Karthikeyan, S., Jacob, G. and Manjunath, B. (20110 Malware images: Visualization and automatic classification. Proceedingsof 8th International Symposium on Visualization for Cyber Security, Article No. 4.
3. Kong, D. and Yan, G. (2013) Discriminant Malware Distance Learning on Structural Information for Automated Malware Classification. Proceedings of the ACM SIGMETRICS/ International Conference on Measurement and modeling of Computer Systems, 347-348.
4. Tian, R., Batten, L. and Versteeg, S. (2008) Function Length as a Tool for Malware Classification. Proceedings of the 3rd International Conference on Malicious and Unwanted Software, Fairfax, 7-8 October 2008, 57-64
5. Siddiqui, M., Wang, M. C. and Lee, J. (2009) Detecting Internet Worms Using Data Mining Techniques. Journal of Systematics, Cybernetics and Informatics, **6,** 48-53.
6. Rieck, K., Trinius, P., Willems, C. and Holz, T. (2011) Automatic Analysis of Malware Behaviour Using Machine Learning. Journal of Computer Security, 19, 639-668.
7. Bayer, U., Comparetti, P. M., Hlauschek, C. and Kruegel, C. (2009) Scablable, Behaviour-Based Malware Clustering. Proceedings of the 16th Annual Network and Distributed System Security Symposium.
8. Anubis. http://sandbox.norman.no
9. Tian, R., Islam, M. R. Batten, L. and Versteeg, S. (2010) Differentiating Malware from Cleanwares Using Behavioural Analysis. Proceedings of 5th International Conference on Malicious and Unwanted Software (Malware), Nancy 19-20 October 2010, 23-30.
10. Biley, M., Oberheid, J., Andersen, J., Morley Mao, Z., Jahanian, F. and Nazario, J. (2007) Automated Classification and analysis of Internet Malware. Proceedings of the 10th International Conference on Recent Advances in Intrusion Detection, 4637, 178-197.
11. Santos, I., Devesa, J., Brezo, F., Nieves, J. and Bringas, P. G. (2013) OPEM: A Static -Dynamic Approach for Machine Learning Based Malware Detection. Proceedings of International Conference CISIS' 12-ICEUTE" 12, Special season Advances in Intelligent System and Computing, **189**, 271-280
12. Islam, R., Tian, R., Battenb, L. and Versteeg, S. (2013) Classification of Malware Based on Intergrated Static and Dynamic Features. Journal of Network and Computer Application, **36**, 646-556.
13. Anderson, B., Storlie, C. and Lane, T. (2012) Improving Malware Classification: Bridging the Static/Dynamic Gap. Proceedings of 5th ACM Workshop on Security and Artificial Intelligence (AISec), 3-14.