A Literature Review of Malware Detection and Classification

A Literature Review

Presented to

Prof. Sharon Stranahan

Department of Computer Science San José State University

In Partial Fulfilments

Of the Requirements for the Class

CS200W

By

Karan Shashin Shah

November 2019

ABSTRACT

Detecting and classifying malware accurately has been one of the biggest topics of research. This review aims to study reasons to detect and classify malware, conventional malware detection techniques, malware detection and classification techniques using machine learning algorithms. In this review, categories of conventional malware detection techniques and malware detection techniques using machine learning is shown. Various machine learning algorithms applied to detect and classify malware are discussed. Although machine learning has improved detecting and classifying malware drastically, we concluded that there is scope of improving accuracy by building efficient machine learning model.

Index term: malware, machine learning, conventional malware detection, static malware detection, dynamic malware detection, hybrid detection.

TABLE OF CONTENTS

I. INTRODUCTION

Malware is a software that corrupts and gains unauthorized access to computer system. It is one of the biggest threat to computer system as well as your privacy.

The first malware called Brian was discovered in 1986 . Virus attacked the floppy boot sector which infected every floppy inserted. Since then thousands of malwares have been developed. Driving factor behind developing a malware has changed to financial gain. Various activities like spam e-mail, web frauds, credit card frauds etc. are performed using malware. As per report published by famous antimalware company called McAfee, malware is used to perform cybercrimes costing up to $500 billion annually.

Lots of research techniques have been developed to detect malware and classify that malware into corresponding malware family. All research techniques focus on improving accuracy of detecting malware. Detection techniques have been broadly categorized into two types

1) Conventional malware detection
2) Malware detection using machine learning

This research review focuses on above mentioned classification of malware detection. Each classification type, various techniques applied to detect malware, their accuracy and limitations are discussed in depth.

Remaining literature review has following organization.

Section II. consists of conventional malware detection. Section III. represents malware detection using machine learning. Section IV. mentions about malware classification. Section V. concludes research review.

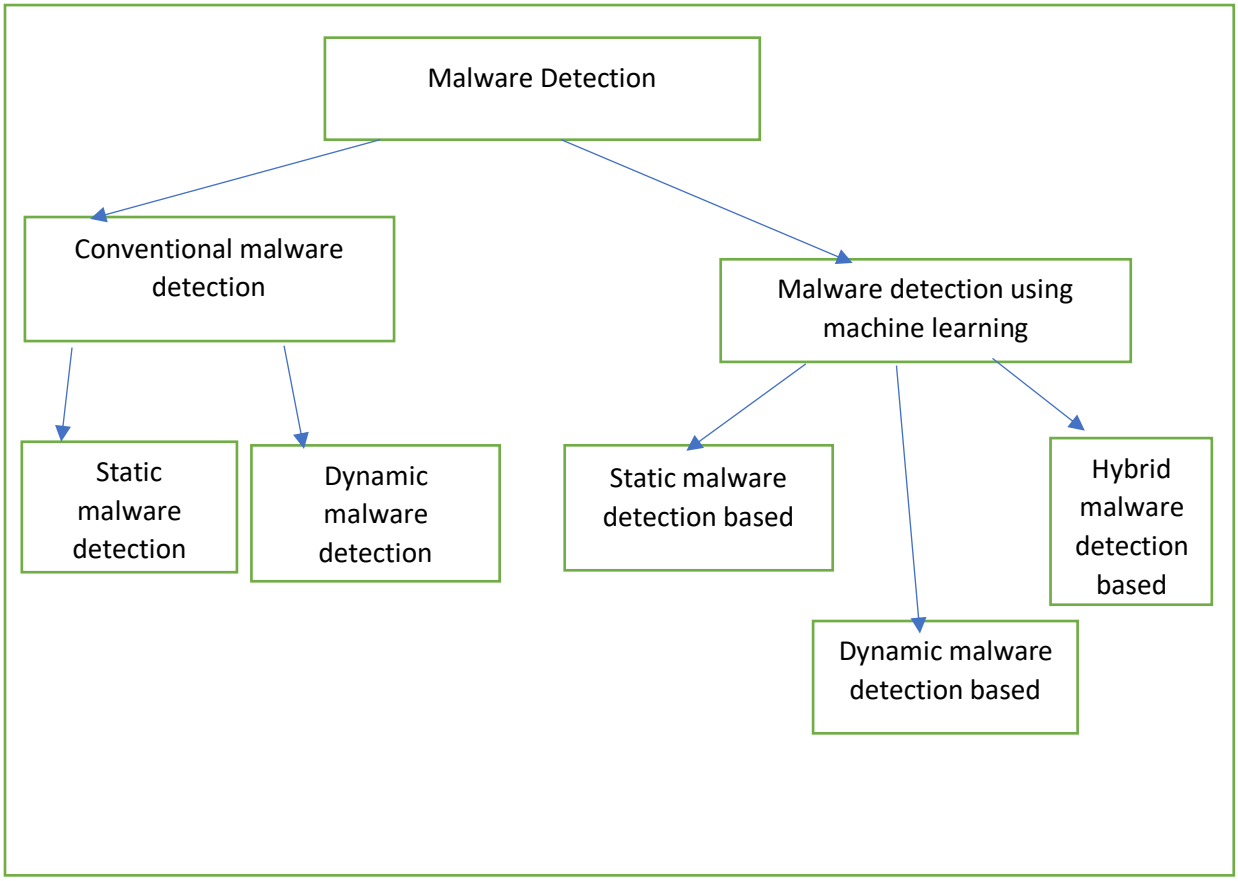

Fig 1. Classification of malware detection

II. CONVENTIONAL MALWARE DETECTION

Conventionally, malware detection technique is divided into two parts which are static malware detection and dynamic malware detection.

Static malware detection is done using the signature test. Signature is unique identification of each malware. These signatures can be opcode, n-grams, malware binaries etc. Here signature is extracted and compared with signature of already detected malwares from the database. If signature matches with any signature from the database then it is classified as malware else it is classified as benign. This method of detecting malware has become obsolete. Attackers are creating dynamic signatures. Dynamic signatures are created by various obfuscation techniques. One of such technique is changing sequence of opcode and include JUMP statement in the code. This method completely changes the opcode making it impossible to detect with signature test. Other shortcomings of static malware detection is mentioned in [1].

Apart from static malware detection, dynamic malware detection technique has been developed. Here, malware is run first in virtual machine or emulated environment. Dynamic malware detection model examines various behaviour of the software. Generally, it observes API calls made by software and classifies it as either malware or benign. Major shortcoming of this technique is that it can only detect malware running in user mode. Malware that run at kernel level, call functions without passing through system call. Hence, this technique cannot be used to detect those malwares. There are various limitations of this method which are mentioned in [1].

III. MALWARE DETECTION USING MACHINE LEARNING

Malware detection using machine learning is categorized into 3 parts

1) Static detection based
2) Dynamic detection based
3) Hybrid detection based

A. Static detection based

In this malware detection technique, machine learning is applied to detect malware using static features like opcode, n-grams, malware binaries etc. This technique uses same features which are used in conventional static malware detection. Using malware signatures, various machine learning models are trained by researchers and accuracy is measured. Schultz [2] trained machine learning model: Naïve Bayes for malware detection using three static features: Dynamic Link Library [DLL], imported and functions referred, strings and byte sequences. Date set of 4266 files were used out which 3265 were malware and 1001 were benign files. Accuracy of 97.11% was achieved. Tian [3] used feature – function length and frequency of function length. Function length is number of bytes code. After training model, they performed k-fold cross validation and achieved 88% accuracy. Siddiqui [4] used variable length instruction as a feature. They did feature reduction using Support Vector Machine [SVM]. They trained Random Forest and decision tree machine learning models for classification and achieved 96% accuracy. Leder [5] performed Value Set Analysis method for malware classification. Value set is extracted from each file . Two value sets are prepared. One set of malware and other set of benign. These sets are compared for classification. Rad [6] used histogram of opcodes for malware classification. They extracted opcodes and made histogram of each opcode instruction. First, they extracted opcodes of malware files and made histogram of each opcode instruction using average of histograms formed by each malware file. In similar way, they made histogram for benign files. After that they tested each file, calculating Euclidean distance from both the sets. They tested data for 100 files only and achieved 100% accuracy. Data set used here is small. This technique starts giving less accuracy as size of data increases. Lakhotia [7] used two features: N-perm feature vector and Term frequency. They weighted these two features using Inverse Document Frequency (IDF) and trained the Nearest neighbour search algorithm. They used dataset of 100 files . They split 40 files for training and 60 files for testing and achieved 86.44% accuracy. Unlike conventional static malware detection techniques which compare signatures from database, machine learning based static malware detection trained machine learning model using signatures. Hence machine learning was able to detect malware

accurately whose signatures are changed by applying obfuscation techniques. So, machine learning based static malware detection is able to overcome shortcoming of conventional static malware detection.

B. Dynamic detection based

In this malware detection method, machine learning is applied on conventional dynamic malware detection techniques. Dynamic features are generated by executing malware in virtual machine. Those dynamic features are used to train machine learning models. Biley [8] executed malware in virtual machine. Report is generated by virtual machine which consists study of various behaviour of malware like files writes, processes creation and network activities. Clustering machine learning model is applied on that using normalized compression distance (NCD) as a distance metric. Bayer [9] executed malware in virtual machine. Behavioural parameters like objects of operating system (OS) and operations of OS are collected. Clustering is applied on that using Local Sensitive hashing (LSH). Using this, they clustered 75,000 samples in less than 3 hours. Rafique [10] collected network traffic logs by executing malware in virtual machine. They extracted feature vector from network traffic logs. This feature vector consists of parameters like msg,fid, rid, proto, dport , sport, sip, dip, dsize, ptreei and endpoint. They used a clustering tool called FIRMA and clustered 16000 samples. Canzanese [11] collected behavioural feature: performance monitor, system call and system call sequence after executing malware in sensors. They applied decision trees and random forest algorithm on 800 malware files and achieved accuracy of 99%. Machine learning based dynamic malware detection is able to overcome shortcomings of conventional dynamic malware detection up to certain extent. Machine learning based dynamic malware can not detect kernel based malware. This method has increased accuracy of detecting malware compared to conventional dynamic malware detection.

C. Hybrid detection based

Ceasare [13] presented method of malware detection which uses both static and dynamic method. Initially, malware is executed inside sandbox to extract opcodes and their frequency. After extracting features, various machine learning algorithms like Bayesian network Decision tree, Support Vector Machine and K-nearest neighbour are applied on the data set of 1000 malware and 1000 benign files. This is the latest proposed method of detecting malware. Very few machine learning techniques have been applied using this method.

IV. MALWARE CLASSIFICATION

Malware is classified into different malware families based on how it corrupts computer system. Major malware families are worm, virus, trojan horse, spyware, bot, rootkit etc. After accurately detecting malware, research has been going on classifying those malwares into corresponding malware family. Microsoft announced malware classification challenge in 2015 on keggle (competitive platform for Data Science). It published dataset of 0.5 Terabytes which consists of disassembly and byte code of more than 20K malware samples.

These malware samples consist of 9 different malware families. This dataset has become benchmark for research and training of machine learning models. This dataset has been cited in more than 50 research papers.

Latest proposal is given by Nataraj [14]. They converted grey-scale image of malware samples and formed vector of 8 bits unsinged integer and labelled them with corresponding malware family. They were able to achieve 98% classification accuracy.

V. CONCLUSION

Malware detection and classification have huge scope of research. Applying machine learning has improved detecting malware drastically compared to conventional malware detecting technique. Now research is focused on improving accuracy of various machine learning models using different features. Developing hybrid-based machine learning model to detect malware is new emerging method. There is need to explore the potential for efficient machine learning model which can show high accuracy.

REFERENCES

[1] M. Andreas, K. Christopher, K. Engin, “Limits of static analysis for malware detection”, 23rd Annual Computer Security Applications Conference, Miami Beach, FL, 2007.

[2] M. G. Schultz, E. Eskin, E. Zadok, S. J. Stolfo, “Data mining methods for detection of new malicious executables”, IEEE Symposium on Security and Privacy, 2001.

[3] R. Tian, L. M. Batten, S. Versteeg, “Function length as a tool for malware classification”, 3rd Int. Conf. on Malicious and Unwanted Software, IEEE, 2008, pp. 69- 76

[4] M. Siddiqui, M. C. Wang, J. Lee, “Detecting internet worms using data mining techniques”, Journ. of Systemics, Cybernetics and Informatics, Vol. 6, No. 6, pp. 48-53, 2008.

[5] F. Leder, B. Steinbock, P. Martini,“Classification and Detection of Metamorphic Malware using Value Set Analysis”, in 4th International Conference on Malicious and Unwanted Software (MALWARE), 2009.

[6] B. B. Rad, M. Masrom, S. Ibrahim,“Opcodes histogram for classifying metamorphic portable executables malware”, In e-Learning and e-Technologies in Education (ICEEE), 2012 Int. Conf. on, 2012.

[7] A. Lakhotia, A. Walenstein, C. Miles, A. Singh,“VILO: a rapid learning nearest-neighbor classifier for malware triage”, Journ. of Computer Virology and Hacking Techniques, Vol. 9, No. 3, pp. 109-123, 2013.

[8] M. Bailey, J. Oberheide, J. Andersen, Z. M. Mao, F. Jahanian, J. Nazario, “Automated classification and analysis of internet malware”, In Recent Advances in Intrusion Detection, Springer Berlin Heidelberg, 2007, pp. 178-197

[9] U. Bayer, P. M. Comparetti, C. Hlauschek, C. Kruegel, E. Kirda,“Scalable, behavior-based malware clustering”, In Network and Distributed System Security Symposium (NDSS), 2009.

[10] M. Z. Rafique, J. Caballero, “Firma: malware clustering and network signature generation with mixed network behaviors,” in Research in Attacks, Intrusions, and Defenses, Springer Berlin Heidelberg, 2013, pp. 144-163.

[11] R. Canzanese, M. Kam, S. Mancoridis, “Toward an automatic, online behavioral malware classification system,” in IEEE 7th International Conference on Self-Adaptive and SelfOrganizing Systems (SASO), 2013, 2013.

[12] M. Z. Rafique, J. Caballero, “Firma: Malware clustering and network signature generation with mixed network behaviors,” in Research in Attacks, Intrusions, and Defenses, Springer Berlin Heidelberg, 2013, pp. 144-163.

[13] S. Cesare, Y. Xiang, W. Zhou,“Malwise – An effective and efficient classification system for packed and polymorphic malware”, IEEE Transactions on Computers, Vol. 62, No. 6, pp. 1193-1206, 2013.

[14] L. Nataraj, S. Karthikeyan, G. Jacob and B. Manjunath, “Malware images: visualization and automatic classification,” in Proceedings of the 8th International Symposium on Visualization for Cyber Security, 2011.