International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

SURVEY ON MALWARE DETECTION USING DEEP LEARNING METHODS

Nourin N. S[1], Sulphikar A[2]

[1]*Student, Dept. of Computer Science and Engineering, LBSITW, Kerala, India*
[2]*Professor, Dept. of Computer Science and Engineering, LBSITW, Kerala, India*

Abstract - *Malware has become one of trending issue nowadays. Malware attackers inject malicious codes into the source code thereby malicious actions take place. Malware can be in different forms (such as trojan-horses, worms, computer viruses etc.). Malware identification continues to be a challenge as attackers comp up with new techniques that tackles the detection methods. In earlier times signature-based detection becomes a milestone for detecting malware. Later on, it becomes inefficient in the present condition due to the speedy increase in the number and versions of malware. Many machine learning techniques are relevant nowadays. In this paper a detailed summary done on different machine learning methods in the area of malware detection which includes Support Vector Machine (SVM), Random Forest, Decision Tree..*

Key Words: *Malware Detection, Machine Learning, SVM, Random forest, Decision Tree.*

1. INTRODUCTION

To cope up with the rapid growth of malware in the internet, malware identification needs to be more efficient. By the origin of ani-malware system it can take counter measures against those not known malwares which are newly generated by the attackers. To make the identification of new malware versions easier, one of the major steps taken by the anti-malware vendors are constantly updating the signature database. Thus, when a new malware which is unknown arises the anti-malware team first crosscheck whether they intercept with those signatures in the database. This makes identification easier and more efficient.

According to AV-TEST [1], total number of known malwares became one billion above samples due to the tremendous generation of malware by attackers. This development wouldn't have been possible without the vigor exhibited by malware authors in the fall of 2019. Detection of new malware samples in June is around 8.5 million samples and in case of July it turns to 9.56 million samples and then jump up above 13 million in the month of august. This monthly rate of detection has not move unsteadily. After culminate in September with a range of 17.70 million, it's actually remained above 15 million with the exception of October at 13.52 million samples. That means this data has proven the dramatical increase in the malware samples.

Earlier times the malware detection has developed with the static and dynamic methods. In the static analysis method, the malware has analyzed statically without actually running it. Only analysis part take place after looking into their structure. But in the case of dynamic analysis the analysis is done by running the malicious code in a supervised environment. Thus, the limitations in the static method can be overcome by those dynamic methods. Later on, hybrid varieties of those static and dynamic approach came which integrate or combines both advantages of static and dynamic methods. Finally, machine learning methods came to exist which makes a remarkable performance in the malware detection field.

In the case of machine learning approaches many techniques like Ripper, Decision Tree, Random Forest, SVM came to exist. Here we are mainly considering with the 3 approaches and their comparison which include Naive Bayes, Random Forest, SVM.

Kephart et al. [2] were first proposed a method which uses machine learning technique for malware detection. Machine learning based method mainly deals with feature extraction and learning model. Feature extraction can be done using many methods includes static features (like n-gram, entropy histogram). In the case of leaning, which makes the model to be trained with the help if features that are extracted.

2. RELATED WORKS

We all know that due to the tremendous increment in releasing new malware samples threaten users' privacies. So, the identification of malware become an inevitable part. Therefore, researches are become more concentrated in this field. For the classification of malware machine learning techniques are widely used. In this related works we are covering major approaches that had done in the malware identification area.

In Schultz et al. [3] used different types of features in their proposal. They used mainly three types of features which include PE headers, strings n-grams, and byte-sequence. These three features are used in three different machine learning algorithms. In Inductive rule-based model, RIPPER is used as classifier and PE headers as features. This technique is based on the list of DLLs (dynamic link library) of 83.62%, the list of DLL function calls (89.36%), DLLs with counted function calls (89.07%). In the case of Probability-based model here uses Naive Bayes as classifier and string n-grams as features (97.11%). In the case of Multi Naïve Bayes model there used byte n-grams (96.88%).

Kolter and Maloof et al. [4] illustrated a boosted decision tree and SVM in their method. Where features are extracted in terms of byte n-grams. They choose n-grams of $n = 4$, byte-

sequence which produces 256 million features. Cross-validation of 10-folds technique helps in the training and testing phase. Binary classification is used to detect malware or benign files. On the other hand, malware identified on the basis of group is known as multi-class classification.

Siddiqui et al. [5] used a feature type of Opcode frequencies as features in their model. The three main classifiers used for malware detection was logistic regression, neural networks and decision tree. Logistic regression having a detection rate of 95% and the neural network is having a detection rate of 97.60% and finally decision tree is having a highest detection rate among the other two that is of 98.40%.

In Tabish et al. [6] uses several machine learning techniques to identify malware files. They stated that their techniques can appropriately distinguish malware from benign despite of its obfuscation using multi-class classification technique. The newness of their method is that the ability to identify obfuscated and packed malware. The main problem in identifying obfuscated malware lies in the obscureness of the structure of the malware file. The attackers will re-write the code of the source file to make it as a malware file intentionally, thus makes it difficult to be found out by anti-malware software.

Gavrilut et al. [7] proposed a method to classify the files as benign or malicious using online Supervised Vector Machine. In online techniques learning take place when the data will pass through the algorithm in a stream. Thereby the model predicts the result by the evaluation of data which is passed and will compare with the actual result. Here they used a combination of 308 features. Learning is done with the help of mistakes that had occurred during the process.

3. ANALYSIS BASED ON MACHINE LEARNING TECHNIQUES

To develop an effective malware identification technique, malware analysis is become an inevitable one. Here it analyzes the functionality and objective of the malicious software. The main aim of the malware analysis is that to interpret the working of malicious codes and thereby create the defenses mechanism against these malicious attacks. There are many approaches have been proposed for the analysis of malwares. In this paper we are going through three different techniques (using SVM, Random Forest, Decision Tree).

3.1. Analysis using SVM

Traditionally, malware identification is carried out using signature method, latterly it became difficult to identify not known samples of malware which are newly released. Finally, machine learning techniques came to exist. One of the best machine learning technique used for the malware identification is linear SVM. Typical SVM algorithms are not suitable for the large amount of data to classify accurately, but in case of linear SVM they classify correctly in the case of large volume of data's too.

In tugsSanjaa and Chuluun et al. [8] they have used datamining approach for the detection and experimentally evaluate malware detection using linear SVM algorithm. Here they gave more importance to two things that is dataset creation and linear data algorithm. SVM is mainly applicable to the data classification which involves training and testing of dataset. SVM create a predicting model for the classification with the help of features. The features are extracted from the testing dataset. Dataset used here is taken from the VXHeaven which contains 271094 malicious executables. They have created 52803 elements of dataset contains 51243 unpacked malicious files and 1560 benign files from different sources. Then they extracted text contents from the dataset to construct vectors.

For the convenient experiment they have splitted the dataset into two where 67% is used as training set and 33% is used as testing test. These splitting is done on the basis of random selection from the original dataset. As a conclusion of their work they showed that linear SVM algorithm for malware identification is beneficial. Average detection rate they got is about 75% and thus proves linear SVM in the case of malware detection gives detection accuracy ranging between 74- 83%.

3.2. Analysis using Random Forest

Random Forest algorithm is also used as a best classifier algorithm for malware. Nowadays, the operating system which is most widely used one is android. Thereby increased in number of cases that malware attackers will target this android operating system. Android users doesn't know whether the user installing apps are trustworthy or not.

In Sethupathi et al. [9] proposed an automated malware detection system named as Maldroid, which is a program that found out whether the application is malicious one or not. This program was evaluated with the large dataset that contains benign apps and malicious apps. Dataset used here for training model contains benign apps and malware apps from the real world. Thus, the dataset helps the model to predict correctly whether the user installing app is a benign app or malicious app. The benign app dataset contains around 30,000 apps where the malicious app dataset contains around 15,000 apps.

Moving to the training phase the generated dataset is used to train the model with given parameters using random forest algorithm. Data Prepossessing is done with the data which helps in the organization of data. The relation between these datasets must be found to extract features. Data cleaning helps to find out data with missing features and values.

The feature extraction is important for prediction and selection. With the collected malicious and benign apps, the features are taken from source codes of decompiled files. The installation package of Android apps is .apk file, that can

be decompiled using Apktool. Experimental outcome proves that Maldroid is capable of identifying malware with comparatively high F1 score of 98%.

3.3 Analysis using Decision Tree

Due to the tremendous increase in the malware types, it is more important to classify the unknown malwares accurately into its own family. Traditionally many antivirus vendors use signature-based method which identify malware based on single feature. The main drawback of signature-based method is that it can only identify versions of malware that have been previously identified. When an unknown malware is released it is failed to detect, this can be overcome by an improved Decision Tree algorithm which can classify malware correctly.

In Sari et al. [10] proposed an improved Decision Tree algorithm that helps in the correct classification of malware. Here they mainly aim to detect the characteristics and structure of existing malware, then classify the malware with the help of Decision Tree technique and then they develop a dashboard using the data that had been analyzed.

In the preprocessing stage they used a Cuckoo Sandbox. It will test the malware samples and thereby generate the report which will determines the malware behavior. As a process of feature extraction and classification, a combining matrix that includes successful Application Program Interfaces, failed APIs and their return code is used. Here they first they form shadow copies of features and the train a Random Forest on the new dataset and finally, classify all features as selected or rejected.

In the visualization part the data's which are classified into family will be visualized by Tableau (visualizing tool). As a result of their experiment they have got 93.3% classification accuracy on multiclass and 94.6% for the binary class. One of the limitations of this method is that they have faced more time consumptions. To overcome these limitations, need to use more combination approaches in machine learning for the classification of malwares.

4. CONCLUSION

Day by day the increase in the number of new malware samples had occurred. To control these growths, we need to identify the malware which are unknown. Many techniques have been come up to control these malware attacks. From that the three machine learning analysis (using SVM, Random Forest, Decision Tree) are discussed here. Also shows the development from traditional signature-based method to machine learning methods.

REFERENCES

[1] https://www.av-test.org/en/statistics/malware/

[2] Jeffrey O Kephart, Gregory B Sorkin, William C Arnold, David M Chess, Gerald J Tesauro, Steve R White, and TJ Watson. 1995. Biologically inspired defenses against computer viruses. In IJCAI (1). 985–996.

[3] Schultz, M.G. et al., 2001. Data mining methods for detection of new malicious executables. In Security and Privacy, 2001. S&P 2001. Proceedings. 2001 IEEE Symposium on. pp. 38–49. Available at: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=924286.

[4] Kolter, J.Z. & Maloof, M.A., 2006. Learning to Detect and Classify Malicious Executables in the Wild. J. Mach. Learn. Res., 7, pp.2721–2744.

[5] Siddiqui, M.., Wang, M.C.. & Lee, J.., 2008. Data mining methods for malware detection using instruction sequences. In Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, AIA 2008. pp. 358–363. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-62849117735&partnerID=40&md5=c8e3822353c1af762c3d350e011103f7.

[6] Tabish, S.M., Shafiq, M.Z. & Farooq, M., 2009. Malware Detection using Statistical Analysis of Byte-Level File Content Categories and Subject Descriptors. Csi-Kdd, pp.23–31. Available at: http://portal.acm.org/citation.cfm?doid=1599272.1599278 [Accessed March 4, 2016].

[7] Gavriluț, D. et al., 2009. Malware detection using machine learning. In Computer Science and Information Technology, 2009. IMCSIT'09. International Multiconference on. IEEE, pp. 735–741.

[8] B. Sanjaa and E. Chuluun, "Malware detection using linear SVM," Ifost, Ulaanbaatar, 2013, pp. 136-138.

[9] Gowtham Sethupathi, Swapnil Siddharth, Vikash Kumar, Pratyush Kumar, Ashwani Yada, "Maldroid: Dynamic Malware Detection using Random Forest Algorithm" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-6, April 2019

[10] Mohd Shaiful Anuar Bin Mohamad Sari, Mohd Aizaini Maarof, "Classification of Malware Family Using Decision Tree Algorithm " UTM Computing Proceedings Innovations in Computing Technology and Applications Volume 2 | Year: 2017 | ISBN: 978-967-0194-95-0.

BIOGRAPHIES

Nourin N.S is pursuing (4th Semester) Master's Degree in Computer Science and Engineering from LBS Institute of Technology for Women, Kerala, India affiliated under Kerala Technical University.

Sulphikar A currently, he is working as an Associate Professor in Computer Science and Engineering, LBS Institute of Technology for Women, Trivandrum, India, affiliated under Kerala Technical University.