MALWARE ANALYSIS USING MACHINE LEARNING

Naga V Vara Prasad Mella [1], P.Maheswari [2]

[1] Assistant Professor **MCA, DEPT, Dantuluri Narayana Raju College, Bhimavaram, Andharapradesh**

[2]**PG Student of MCA, Dantuluri Narayana Raju College, Bhimavaram, Andharapradesh**

Abstract Malware analysis forms a critical component of cyber defense mechanism. In the last decade, lot of research has been done, using machine learning methods on both static as well as dynamic analysis. Since the aim and objective of malware developers have changed from just for fame to political espionage or financial gain, the malware is also getting evolved in its form, and infection methods. One of the latest forms of malware is known as targeted malware, on which not much research has happened. Targeted malware, which is a superset of Advanced Persistent Threat (APT), is growing in its volume and complexity in recent years. Targeted Cyber attack (through targeted malware) plays an increasingly malicious role in disrupting the online social and financial systems. APTs are designed to steal corporate / national secrets and/or harm national/corporate interests. It is difficult to recognize targeted malware by antivirus, IDS, IPS and custom malware detection tools. Attackers leverage compelling social engineering techniques along with one or more zero day vulnerabilities for deploying APTs. Along with these, the recent introduction of Crypto locker and Ransom ware pose serious threats to organizations/nations as well as individuals. In this paper, we compare various machine-learning techniques used for analyzing malwares, focusing on static analysis

Index Term: - cyber defence, cyber-attacks, crypto locker,randomware,APTS

I Introduction

"Malware" is an abbreviation for "malicious software", it is used as a single term to refer to Viruses, Trojans, Worms, etc. These programs have a variety of features, such as stealing, encrypting or deleting sensitive data, modifying or hijacking basic computer functions, and monitoring computer activity. Show user permission.

1.1 Types of Malware:

- Computer virus It is generally a program that is installed outside the user's will and can cause damage to both the operating system and the hardware (physical) elements of a computer. Effects generated by the virus: Slowing down the computer's working speed until it crashes Worms Computer worms are programs with destructive effects that use communication between computers to spread. Worms have common features with viruses, ii. Worms are able to multiply like viruses, but not locally, but on other computers. I use computer networks to spread to other systems. It can hide certain files (usually their own) in case an antivirus program scans Spyware is a category of cyber threats, which describes malware created to infect PC systems and then initiate illegal activities. these. In most cases, the functionality of these threats depends on the intentions of their vendors: some parts of spyware threats can be used to collect personal information (login names, passwords, and other personally identifiable data) and send them to their owners via hidden internet connections, while other spyware viruses can track their victims and collect information about their browsing habits. These are used to track people and record their most visited websites as well as the actions taken when they were visited. This information is generally used by various third parties for marketing and promotional purposes, so spyware can also lead to an increase in the number of spasm.

Malware analysis is necessary for the development of effective techniques for detecting infested files.

This analysis is the process of observing the purpose and functionality of a malware program. There are 3 analysis techniques that have the same purpose: to explain how a malware works and what its effects are on the system, but the time and knowledge required are very different.

1.2 Static analysis

It is also called code analysis. That is, the malware software code is observed to gain knowledge about the operation of malware functions. This reverse engineering technique is performed using disassembly, decompilation, debugging, and source code analysis tools. We will be following this technique as is that it is free from the overhead of execution time.

1.3 Dynamic analysis

It is also called behavioral analysis. Infected files are analyzed during execution in an isolated environment such as a virtual machine, simulator, or emulator. After the execution of the file, the behavior and its effects on the system are monitored.

1.4 Hybrid analysis

This technique is proposed to overcome the limitations of static and dynamic analysis. First, it analyses the specification of the signature for any malware code and then combines it with the other behavioral parameters to improve the complete analysis of malware. Due to this approach, hybrid scanning exceeds the limits of static and dynamic scans

2 Literature survey

You Only Look Once: Unified, Real-Time Object Detection, by Joseph Redmon. Their prior work is on detecting objects using a regression algorithm. To get high accuracy and good predictions they have proposed YOLO algorithm in this paper [1]. Understanding of Object Detection Based on CNN Family and YOLO, by Juan Du. In this paper, they generally explained about the object detection families like CNN, R-CNN and compared their efficiency and introduced YOLO algorithm to increase the efficiency. Learning to Localize Objects with Structured Output Regression, by Matthew B. Blaschko. This paper is about Object Localization. In this, they used the Bounding box method for localization of the objects to overcome the drawbacks of the sliding window method.

2.1. Viruses , Trojans , And Spyware , Oh My! The yellow brick road to coverage in the land of internet OZ

Author: Roberta D. Anderson

Tort Trial & Insurance Practice Law Journal

Abstract

Every company is at cyber risk. The headlines confirm the reality: cyber attacks are on the rise with unprecedented frequency, sophistication, and scale. And they are pervasive across industries and geographical boundaries. As serious cyber threats are making daily headlines, regulations surrounding data privacy and security are proliferating. With data security breaches, denial of service, and other attacks and loss of data on the rise, addressing and mitigating cyber risk is a top priority among companies across the globe. It is abundantly clear that network security alone cannot entirely address the issue of cyber risk; no firewall is unbreachable, no security system impenetrable. Insurance can play a vital role in a company's overall strategy to address, mitigate, and maximize protection against cyber risk. This fact has the attention of the Securities and Exchange Commission. In the wake of "more frequent and severe cyber incidents," the SEC's Division of Corporation Finance has issued guidance on cyber security disclosures under the federal securities laws. The guidance advises that companies "should review, on an ongoing basis, the adequacy of their disclosure relating to cyber security risks and cyber incidents" and that "appropriate disclosures may include" a "[d] description of relevant insurance coverage."

2.2 An improved Android malware detection scheme based on an evolving hybrid neuro-fuzzy classifier (EHNFC) and permission-based features

Author: Alter Alta her Tasha

Abstract

The increasing number of Android devices and users has been attracting the attention of different types of attackers. Malware authors create new versions of malware from previous ones by implementing code obfuscation techniques. Obfuscated malware is potentially contributed to the exponential increase in the number of generated malware variants. Detection of obfuscated malware is a continuous challenge because it can easily evade the signature-based malware detectors, and behavior-based detectors are not able to detect them accurately. Therefore, an efficient technique for obfuscated malware detection in Android-based smart phones is needed. In the literature on Android malware classification, few malware detection approaches are designed with the capability of detecting obfuscated malware. However, these malware detection approaches were not equipped with the capacity to improve their performance by learning and evolving their malware detection rules. Based on the concept of evolving soft computing systems, this paper proposes an evolving hybrid neuro-fuzzy classifier (EHNFC) for Android malware classification using permission-based features. The proposed EHNFC not only has the capability of detecting obfuscated malware using fuzzy rules, but can also evolve its structure by learning new malware detection fuzzy rules to improve its detection accuracy when used in detection of more malware applications. To this end, an evolving clustering method for adapting and evolving malware detection fuzzy rules was modified to incorporate an adaptive procedure for updating the radii and centers of clustered permission-based features. This modification to the evolving clustering method enhances cluster convergence and generates rules that are better tailored to the input data, hence improving the classification accuracy of the proposed EHNFC. The experimental results for the proposed EHNFC show that the proposal outperforms several state-of-the-art obfuscated malware classification approaches in terms of false negative rate (0.05) and false positive rate (0.05). The results also demonstrate that the proposal detects the Android malware better than other neuro-fuzzy systems (viz., the adaptive neuro-fuzzy inference system and the dynamic evolving neuro-fuzzy system) in terms of accuracy (90%).

3. Implementation Study

Network security is the key challenges for much enterprise level application which are mainly used identify the security threats associated across different types of attacks by hackers as well as intruders who always involve in spoofing the data during the data transmission.

3.1proposed methodology

This has, in turn, stimulated the allocation of the technological means geared towards the refinement and streamlining operations of Businesses because of performance gains within the competitive scenario of the overall market. The vitality of the approaches underlined within the purview of analytics practices to extract the optimal potential has been duly emphasized. The present framework of Data Analytics practices is extrapolated based on the Data Analytics practices within the context of IT Business infrastructure. Furthermore, the limitations and the future scope of the Data as mentioned above Analytics technologies in light of optimal decision undertaking within organizational context while considering the various risk management approaches fostering enhanced consumer experience and improved innovation and development practices within industries have been duly considered

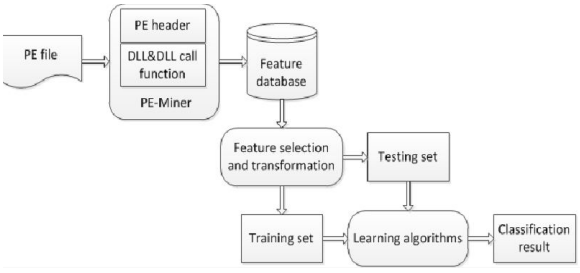

Fig 1: proposed model

3.2 Methodology

3.2.1 DATA PREPROCESSING

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

3.2.2 DATA EXPLORATION

Data exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data.

3.2.3 MODEL CREATION

A machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.

3.2.4 TRAINING AND TESTING

Training data is the initial dataset you use to teach a machine learning application to recognize patterns or perform to your criteria, while testing or validation data is used to evaluate your model's accuracy. You'll need a new dataset to validate the model because it already “knows” the training data.

3.2.5 PREDICTION

Training data is the initial dataset you use to teach a machine learning application to recognize patterns or perform to your criteria, while testing or validation data is used to evaluate your model's accuracy. You'll need a new dataset to validate the model because it already “knows” the training data

4 Results and Evolution Metrics

Fig2:- here it predicts the malware or not

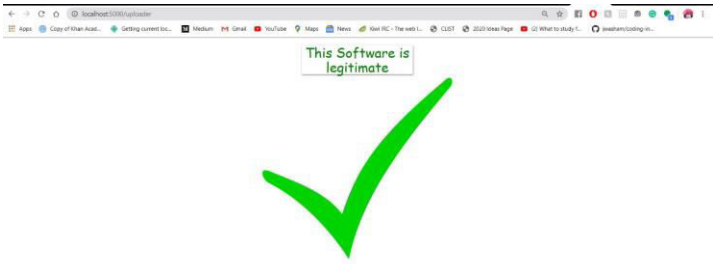

Fig 3 here it predicts it is not a malware

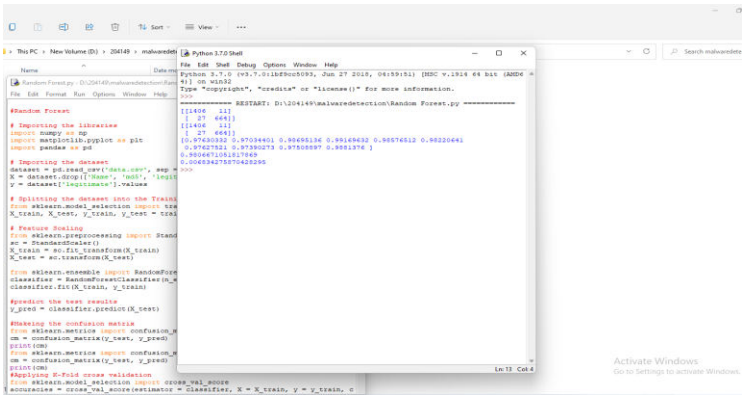

Fig 4: - accuracy of the model

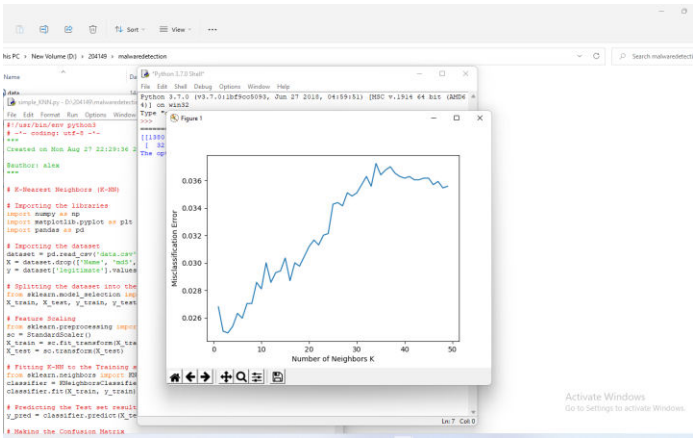

Fig 5:- accuracy graph of the model

5 Conclusion

The aim of this paper is to present a machine learning approach to the malware problem. Due to the sudden growth of malware, we need automatic methods to detect infested files. In the first phase of the work, the data set is created using infested and clean executables, in order to extract the data necessary for the creation of the data set, we used a script created in Python. After creating the data set, it must be ready to train machine learning algorithms. The algorithms used are: decision trees, Random Forest, Naïve Bayes, GradientBoost and ADABoost presented comparatively. After applying the best accuracy algorithms, it had an Random Forest algorithm with an accuracy of 99.406012 %. This work demonstrates that Random Forest is the best algorithm for detecting malicious programs. In the future, this accuracy can be improved, if we add a much larger number of files in the data set to drive the algorithms. Each algorithm has several parameters that can be tested with different values to increase their accuracy. This project can reach the application level with the help of a library called pickle, to save what the algorithm has learned and then we can test a new file to see if it is clean or infected. Static analysis has also proven to be safer and free from the overhead of execution time.

.

6 References

1 Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). Introduction to pathogens.

2 Altaher A (2016) An improved Android malware detection scheme based on an evolving hybrid neuro-fuzzy classifier (EHNFC) and permission-based features. *Neural Comput Appl* 28:4147–4157. https://doi.org/10.1007/s00521-016-2708-7

3 Alzarooni, K. M. A. (2012). *Malware variant detection* (Doctoral dissertation, UCL (University College London)).

4 Anderson, R. D. (2014). Viruses, Trojans, and Spyware, Oh My! The Yellow Brick Road to Coverage in the Land of Internet Oz. *Tort Trial & Insurance Practice Law Journal*, 529-610.

5 Anwar, S., Mohamad Zain, J., Zolkipli, M. F., Inayat, Z., Khan, S., Anthony, B., & Chang, V. (2017). From intrusion detection to an intrusion response system: fundamentals, requirements, and future directions. *Algorithms*, *10*(2), 39.

6 Bat-Erdene M, Park H, Li H, Lee H, Choi MS (2017) Entropy analysis to classify unknown packing algorithms for malware detection. *Int J Inf Secur* 16(3):227–248. https://doi.org/10.1007/s10207-016-0330-4

7 Bayer, U., Kruegel, C., & Kirda, E. (2006). *TTAnalyze: A tool for analyzing malware* (pp. 180-192). na.