MALWARE DETECTION USING HONEYPOT BASED ON MACHINE LEARNING

GK. CHAKRAVARTHY, M. BHARADWAJ PHANI DATTA,

M. MOUNIKA & Y. RAMAKRISHNA

Assistant Professor, Department of Computer Science and Engineering,

Koneru Lakshmaiah Educational Foundation, Andhra Pradesh, India

ABSTRACT

Malware is nothing but a malicious software which is a threat to information present in any system. Virus is also a type of malware which when executed self-replicates and modifies other computer programs by inserting its own code. There are different types of malwares like worms, spyware, trojan horse, logic bombs, ransomware, backdoors, rootkits, keyloggers etc., which attack the system and destroys the data present in it. In 2019 nearly 989,432,403 new malwares were recorded. Most of the malwares are Scripts and HTML malware. Adware is that the most importantly increased malware. Adware malware is the name of a program designed to display advertisements on a computer, which redirects the search request to advertising websites and collects marketing data about that person. To protect our systems from these kinds of malware we use different Security framework gadgets, for example, antivirus, firewall, and IDS. But these security system devices failed to protect the data due to the quick spread of PC malware and the expanding number of marks. Other than signature-based security frameworks it is hard to recognize new techniques, infections or worms utilized by assailants. One other option in distinguishing malware is to utilize honeypot with AI. Honeypot is goes about as a virtual framework which can be utilized as a snare for bundles that are suspected while AI can be utilized to recognize malware by arranging classes with various kinds of AI calculations, for example, K-implies Clustering and Random Forest which are utilized as order calculations in this paper. In this paper, we propose engineering structure as an answer for identify malware and furthermore introduced the compositional proposition and disclosed the exploratory technique to be utilized.

KEYWORDS : Honeypot, Malware; Machine learning, K-means & Random Forest

Received: Apr 13, 2020; **Accepted:** May 04, 2020; **Published:** May 29, 2020; **Paper Id.:** IJCSEITRJUN20205

INTRODUCTION

For whatever purpose the computer may be used it is necessary to protect the data from malware. Previously to protect data we used different techniques like backing up the data into any trusted backup provider by which we can recover the data even if something happens to the system. Another way to protect data is to make sure that the system is up to date. Being aware of the strange emails. Be careful while browsing the internet because clicking on any link may take you to another information which is not related to the required information. Being cautious with social networking spams and removing them. Installing antivirus software and anti-spyware. All these precautions can be taken only if you are aware of the malicious software. But now hackers are getting sneaky and it is becoming more difficult to identify what type of malware is attacking the system. To identify the type of malware and the intention of the attacker honeypots are introduced. Honeypots are the decoy systems that are located within the network to monitor the trails of hackers and alert the system administrator about the intrusion. Basically, honeypot acts as the actual system which is seen by the attacker and he does attack on this honeypot through which the system administrator gets time to identify what type of attack is going to happen and can protect the data from the hacker.

To advance the activity of the honeypots there comes honeypots with machine learning. Here the output given from the honeypots are given to some machine learning algorithm to identify malicious and non-malicious data or to identify particular type of website which are not related to the data or spam emails etc., so that it will be easy for the administrator to separate and block the entry so such malicious data.

According to data published by Kaspersky, number of malwares remains to increase in intensity. In the subsequent half of 2019 there were 717,057,912 attacks launched from online resources around 203 countries. Here 217,843,293 unique URL's are used to trigger a Web base attack. In this there are 228,206 users who are affected by these attacks and used their systems to steal money via online bank accounts. The file-based attacks are around 240,754,063. Here the ransomware attacks are defeated around on 232,292 users, in view of the classification, the most noteworthy number of malwares is in Trojan Horse malware like mobile banking Trojans and mobile ransomware Trojans are come into existence more in this year [1].

By the end of the year 2019 the malware attacks had increased to 989,432,403 and the unique URL's attacks had recognized around 560,025,316 and the attempts to steal money from users around 197,559. Ransomware attacks were defeated around on 229,643 computer users and the file-based attacks around 230,051,054. [2]

The reason for the honeypot is to gain proficiency with the procedures of the apparatus or the technique utilized by the assailant [5]. AI can be utilized to distinguish the nearness of malware. AI works by arranging a class in a characterized dataset. Arrangement is utilized to separate two classes in particular amiable and malware [5].

This paper focus on malware classification using K-means clustering algorithm. Then this malware class will be analyzed using Random Forest algorithm. In view of the examination it is realized that the normal exactness can arrive at 90% with the aftereffects of the assessment demonstrating the most noteworthy productivity results [7][8].

In this paper we propose an engineering for identifying malware utilizing honeypot and AI. we likewise disclosed the techniques to be utilized in leading tests. We use k-means and irregular backwoods calculations. These calculations are demonstrated to be unrivaled as far as exactness, identification rates, bogus admonition frameworks and so on., [9]. The engineering that we propose is intended to have the option to distinguish malware dependent on its conduct. Besides the framework that we fabricate can prepare without anyone else dependent on the consequences of discovery. With the goal that new strategy for malware will be distinguished. This paper is isolated into five areas as follows: Part I, we talk about the presentation; part II, we present papers related investigations; Part III we examine the proposed design; in segment IV, we examine the strategies that will be utilized in leading trials; segment V we present outcomes and research openings later on.

2. RELATEDWORK

PhilippeOwezarski[10]proposed a computerized strategy for abnormalities arrangement and portrayal on the system without oversight. This exploration can identify DoS, DDoS, and Port Scan assaults.

AbdallahGhourabi [11] propose the utilization of a honeypot to identify and contemplate assaults on web administrations. Honeypot is utilized to learn new programmers' procedures, instruments and inspirations to ensure the framework. This exploration can identify SQL Injection assaults, Session Hijacking, EXPATH Injection, and Tampering Parameters.

FatihHaltaş [12] presents a programmed bot location framework that is tainted with BFH (Bot Trough Honeypot Finder). This investigation of the BFH System naturally sends malware got to the arrangement unit to manufacture a family gathering. At that point, tests are consequently given to the preparation unit for displaying and location of NetFlow information.

KaterinaGoseva-Popstojanova [13] characterize malignant assaults on a web meeting. The consequences of this investigation can arrange DoS assaults, Password Cracking, Email Harvesting, Spam, SQL Injection, XSS

LikMuhamad Malik Matin [] propose the architecture and design for distinguishing malware applying honeypots and machine learning using Support Vector Machine and Decision Tree algorithms.

3. PROPOSEDARCHITECTURE

Architecture

In fig.1 given below we show the complete compositional structure of honeypot and AI. In this structure there are organize parts, Routers, Honeypot, Data Analysis, and Real System. In this procedure, all the outer system traffic goes to the inner system. The switch is liable for interfacing the outside system into the interior system and advances the parcels into the honeypot. Honeypot can catch traffic parcels that are gone into the interior system and put away in a fake server. At that point the caught bundles are utilized for investigation purposes.

DataAnalysis

In the information investigation module, learning is done on the malware location framework. Dataset are accommodated preparing and testing. Here we utilized unaided learning with marked dataset. Moreover, the K-implies calculation and Random Forest calculation are utilized in arranging the dataset through which we can create high exactness with the high effective outcomes. The arrangement calculation will attempt to distinguish the class climate it is malware information or favorable information. Malware information is removed and answered to the human specialists for the upgradation of existing programming. Besides, re-preparing is done to improve location exactness.

Figure 1

4. EXPERIMENTALMETHOD

In this experimental technique, we explain the dataset, percentage split, evaluation method.

Dataset

The data that is registered by Honeypot is exported as dataset and it is used for the experiment. This dataset consists of 700002 training samples (22215 malicious samples, 677787 benign samples). The distribution of the dataset can be explained in Figure 2

Figure 2: Distribution of Malicious Data.

Clustering

Clustering is used to partition all the observations into K clusters in which each observation belongs to the cluster to its nearest mean with the goal that each datum will encounter by and large approval. The division of the dataset can be portrayed in Figure 3.

Figure 3: Distribution of Dataset into Cluster.

Random Forest

We use Random Forest Algorithm to evaluate the result of our Clustering algorithm. It is an ensemble learning method; it will generate multitude of decision trees at training time and give output as individual trees. Decision tree for random forest can be described in fig.4. and the figure 5.

Figure 4: Decision Trees in Random Forest.

Figure 5: Output of Random Forest.

Evaluation Method

Estimation of AI precision dependent on disarray lattice can be clarified in the accompanying tab. I:

Table1: ConfusionMatrix

Parameter		Prediction	
		Malware	Benign
Reality	Malware	True Positive	False Positive
	Benign	False Negative	True Negative

a. True Positive(TP)

Malware correctly identified

b. False Negative(FN)

malware that neglected to be recognized. which means bogus negative shows that the grouping results are favorable.

c. True Negative (TN) Benign correctly identified

d. False Positive(FP)

Favorable that neglected to be distinguished. This implies bogus positives show the order aftereffects of malware [11].

The four potential aftereffects of explanation did by AI; it very well may be seen the precision of a technique used to identify malware. There are four parameters to decide the exactness utilized in AI, to be specific True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall and Accuracy as follows:

$$TPR = \frac{TP}{TP+FN} \quad (1)$$

$$FPR = \frac{FP}{FP+TN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Genuine positive Rate (TPR) is the recognized benevolent proportion. The bogus Positive Rate (FPR) is the proportion of misidentified malware. Exactness is importance in recognizing from the aftereffects of explanation given by considerate yield, review is recognition rate or affectability and precision is the proportion that accurately distinguished [11].

CONCLUSIONS

In this paper we initially proposed the architecture where we used honeypots and machine learning algorithm to identify, analyze and classify the data. Data packets from the external network are entered into internal network through routers. Honeypot in the internal network acts as the original system and all the packets enter the honeypot. The output from the honeypot are taken as the dataset which are sent to the machine learning algorithm i.e., k-means and random forest algorithms. These algorithms analyze and classify the data into clusters as malicious and non-malicious data. Here we used

unsupervised learning technique as it provides more accurate and efficient results than supervised learning technique. This classified data can be provided to administrator to improvise further as self-healing of networks by using self-learning techniques.

REFERENCES

1. Kaspersky Security, "Malware Report: Quarterly Report August 2019," [Online]. Available: <https://securelist.com/it-threat-evolution-q2-2019-statistics/92053/>.
2. Kaspersky Security, "Malware Report: Quarterly Report August 2019," [Online]. Available: <https://securelist.com/it-threat-evolution-q2-2019-statistics/95269/>.
3. I. Firdausi, C. lim, A. Erwin, and A. S. Nugroho, "Analysis of Machine learning Techniques Used in Behavior-Based Malware Detection," in *2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies*, Jakarta, Indonesia, 2010, pp. 201–203.
4. B. Nagpal, N. Singh, N. Chauhan, and P. Sharma, "Catch: Comparison and analysis of tools covering honeypots," in *2015 International Conference on Advances in Computer Engineering and Applications*, Ghaziabad, India, 2015, pp. 783–786.
5. U. Pehlivan, N. Baltaci, C. Acarturk, and N. Baykal, "The analysis of feature selection methods and classification algorithms in permission based Android malware detection," in *2014 IEEE Symposium on Computational Intelligence in Cyber Security (CICS)*, Orlando, FL, USA, 2014, pp. 1–8.
6. M. N. Mahrin et al., "MALWARE PREDICTION ALGORITHM: SYSTEMATIC REVIEW," . Vol., no. 14, p. 17, 2005.
7. D.-T. Truong and G. Cheng, "Detecting domain-flux botnet based on DNS traffic features in managed network: Detecting domain-flux botnet," *Security and Communication Networks*, vol. 9, no. 14, pp. 2338–2347, Sep. 2016.
8. Y. Ye, L. Chen, D. Wang, T. Li, Q. Jiang, and M. Zhao, "SBMDS: an interpretable string based malware detection system using SVM ensemble with bagging," *Journal in Computer Virology*, vol. 5, no. 4, pp. 283–293, Nov. 2009.
9. KamarularifinAbdJalil,MuhammadHilmiKamarudin,andMohamadNoormanMasrek, "Comparison of Machine Learning algorithms performance in detecting network intrusion," in *2010 International Conference on Networking and and InformationTechnology*, Manila, Philippines, 2010, pp. 221–226.
10. P. Owezarski, "Unsupervised classification and characterization of honeypot attacks," in *10th International Conference on Network and Service Management (CNSM) and Workshop*, 2014, pp.10–18.
11. IikMuhamad Malik Matin, Budi Rahardjo "Malware Detection Using Honeypot and Machine Learning" in *The 7th International Conference on Cyber and IT Service Management (CITSM 2019)*
12. Segal, Mark "Machine Learning Benchmarks andRandom Forest Regression" 2003 Kluwer Academic Publishers. Printed in the Netherlands.