International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com || Volume 02 - Issue 10 || October 2016 || PP. 45-49

Malware Classification and Machine Learning: A Survey

Devyani Bhamare[1], Poonam Bhamare[2]
[1]*Department of IT, Sanjivani College of Engineering, Kopargaon, India*

Abstract: Malicious software, referred to as malware, is one of the major threats on the Internet today. Due to vast use of Internet it found that many system are infected from malware which in form of computer virus, Trojan Horse, Worms, Rootkit, backdoor, evasion etc. In this paper we reviewed different approaches for malware analysis. Malware Analysis and classification with machine learning techniques is also discussed. This paper aimed to provide introduction about malware and its related issues.
Keywords: Machine Learning, Malware, Malware Analysis , Classification

I. INTRODUCTION

Malware means Malicious Software which is programmed for disrupting or denying operations or stealing private information or to gain unauthorized access to system and other resources. Malware is also used against individuals to gain information such as personal identification numbers or other details, bank credit card numbers, password. Malware are great challenge in our day to day life. According to the Global Threat Report from Cisco Security Intelligence Operations, there were 287,298 "unique malware encounters" in June 2011, double the number of incidents that occurred in March. To help mitigate the threat of malware, researchers at the SEI are investigating the origin of executable software binaries that often take the form of malware [7]. It is found that malware is one of the most terrible and major security threats which found on Internet. Malware are classified in many variations like virus, worms, Spy ware and Botnet and so on. Among these different classes it is found that the one type of malware may have characteristics of which are similar to other type. Malware is great challenge in our day to day life. The malware are continuously growing in volume, variety and velocity. This means that they have becoming more sophisticated and uses new approaches to harm a computers and mobile devices too. Mcfee has given that near about 10,000 new malware samples are introduced every day [1]. Due to new and readily available and easy tools the percentage of cyber threat is increased. If we compare it is found that traditional malware were broad, know, open and one time but the advanced malware are targeted, unknown, and stealthy, personalized. So traditional malware are hiding, replicate and disable the host protection once they go inside of the host and install themselves and call their command and control services for their further instructions like steal data, infect the machine as on [5]. TO identify the inherent similarities and shared pattern between malware using ,machine learning becomes recent field. As Machine learning is filled in which learning means identify patterns. As human uses pattern to recognize the object are color, shape, sound and smell. So similarly machine can identify pattern in sequence of bits of collection of malware. Therefore machine learning has a natural fit with Malware analysis as it can learn and fid patterns in malware.

The paper is organized as following. Firstly we discuss about the Malware. Machine Learning and Malware Analysis is given in section III and section IV respectively. Secttion IV discussed about different machine learning techniques and its related work for malware classification. Finally conclusion is given section VI.

II. MALWARE

Malware means Malicious Software i.e is any software used to disrupt computer operations, gather sensitive information, gain access to private computer systems, or display unwanted advertising. In1990, Yisrael Radai said that, malicious software was referred to as computer viruses. The first category of malware propagation concerns parasitic software fragments that attach themselves to some existing executable content. The fragment may be machine code that infects some existing application, utility, or system program, or even the code used to boot a computer system. Malware is defined by its malicious intent, acting against the requirements of the computer user, and does not include software that causes unintentional harm due to some deficiency [9]. As there no any accepted standards fot classification of malware int h industry but depending on methods of malware propogation they classified in to virus, worm, and trojan (short for Trojan horse). **Virus: It is malware that attaches copy of itself to host** Propagation by infecting removable media was the only method for transmission available prior to the Internet, and this method is still in use today. For instance, modern viruses travel by infecting USB drives. This method is still necessary to reach computer systems that are not connected to the Internet, and is hypothesized as the way Stuxnet was transmitted.

A Trojan Horse: It propagates the same way its name sake entered the city of Troy, by hiding inside something that seems perfectly innocent. The earliest trojan was a game called ANIMAL. This simple game would ask the user a serious of questions and attempt to guess what animal the user was thinking of. When the game was executed, a hidden program, named PERVADE, would install a copy of itself and ANIMAL to every location the user had access to. A common modern example of a trojan is a fake antivirus, a program that purports to be an anti-virus system but in fact is a malware itself.

A worm: It is a self-propagating malware. Whereas a virus, after attaching itself to a program or document, relies on an action from a user to be activated and spread, a worm is capable of spreading between network connected computers all by itself. This is typically accomplished one of two ways: exploiting vulnerabilities on a networked service or through email. The worm CODE RED was an example of the first type of worm. As Malwares are rapidly increasing due to internet , It require automatically detect malwares. Malware Analysis has a typical goal is simply automatically detect malware as soon as possible, remove it, and repair any damage it has done. To accomplish this goal, software running on the system being protected (desktop, laptop, server, mobile device, embedded device, etc.) uses some type of "signatures" to look for malware. When a match is made on a "signature", a removal and repair script is triggered. There are two common technique used for malware detection and classification: Signature Based and behavior based. In signature based techniques uses a common sequence of bytes that appear in the binary code of malware family to detect and identify malware samples. This technique is fast as it does not require executing the code to identify them. It may give inaccurate results. It also requires prior knowledge of signature which associated with the malware families. In Behavior-based technique is depends upon the artifacts which are created during executing. This is quite expensive as it requires running the sample to obtain artifacts and features.

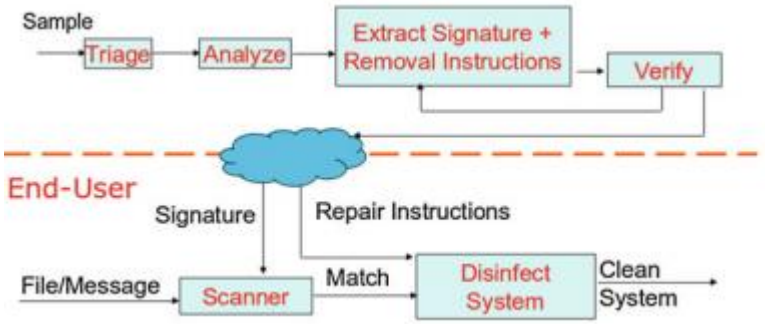

Figure 1: Phases of malware analysis Pipeline [10]

As shown in Fig. 1 Input the malware analysis is suspicious program which need to analyze. Then Signature created and verified. Then that signature is propagated to end system and used to detect or remove or repair malware.

III. MACHINE LEARNING

Machine Learning is subfield of computer science that gives computer ability to lean without being explicitly programmed [9]. Machine learning explores study and construction of an algorithm that can learn from data and makes predictions on data, the algorithms which makes data driven predictions or decisions on data through building a model from sample input.

When pattern are recognized using a "labeled data" it is termed as "Supervised learning" i.e. data is described (labels) and category (class) is known. While on the other hand, if the data is "unlabeled" it is termed as "Unsupervised learning" i.e. the category or class is unknown at the time of learning. Similar to machine learning, other related terms to Pattern recognition are data mining and knowledge discovery in databases as they largely overlap in their scope. Supervised learning, unsupervised learning, semi supervised and reinforcement learning are the categories of machine learning.

In supervised learning the computers are present with examples input and their desired labeled output, which aims to learn a general rule that maps input with output. But in unsupervised no output labels are present, depends on structure of its input which have goal to discover hidden patterns and makes group. By considering desired output of machine learning system Machine learning task also categorized into classification, regression, dimensionality reduction and clustering

IV. MALWARE ANALYSIS

It is required to analyzed the actual intension and risk associated with that malware before creating the signature for newly arrived malware. Many times male ware program and their capability can be checked and observed by executing the code in safe environment or by checking its code. It shows that a majority of malware families generate from the same origin. Depending upon these type of checking malware can be analyzed by of two types static analysis and dynamic analysis.

The Static analysis means analyze the malicious software without executing the code. The Dynamic Analysis means analyze the dynamic behavior means to analyze that how malicious code will interact with the system. This is can be done by executing the code in controlled environment such as simulator, sandbox, emulator. There are various techniques that can be applied to perform dynamic analysis. There are different monitoring tools for different purpose such as for file system and registry monitoring tools are Process Monitor and Capture BAT, for process monitoring tools are used such as Process Explorer and Process Kackerreplace and Wireshark used for network monitoring and so on. So before executing the malware samples such appropriate monitoring tools are installed and activated [3]. There are many automated tools are available for dynamic analysis of malware for example Norman Sandbox, CW sandbox, Anubis. The analysis report of these tool gives in depth understanding of malware behavior and the action performed by them [3].

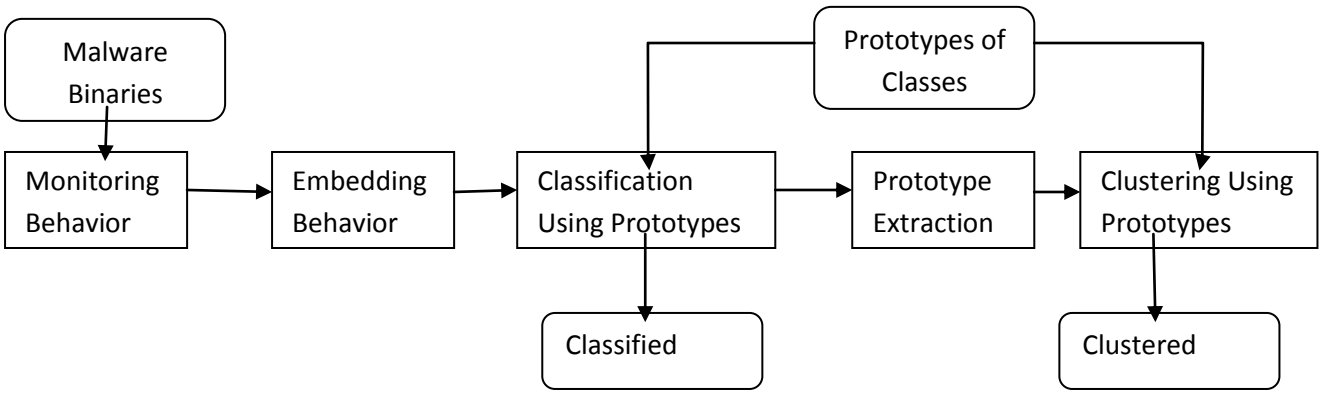

Figure 1. Schematic overview of analysis framework [8].

As shown in figure Fig. 1 show the framework for automatic analysis of malware behavior. The Author[8] has explained framework which allows the automatically identifying malware with similar behavior to form cluster (clustering) and aslo assing unknow malwrae to respective class (classification). In the figure 1 shows that First malware binaries are executed and monitored then the report generated by this process for each binary. Then Machine learning Techniques for clustering and classification are applied on these report to identify the classes of malware [8]. In this author has used two datasets of malware behavior A reference data set which consist of known classes of malware and application dataset from

V. MACHINE LEARNING TECHNIQUES FOR MALWARE ANALYSIS

Machine Learning uses statistical techniques to learn pattern from large datasets. Current methods for malware analysis tools are manual techniques which may take more time and effort for analysis of malware and recently huge work is going on automated techniques to analyze the malware. Machine Learning techniques can be applied to analyze the malware as Machine learning consist of two phase training and testing. So for malware analysis by machine learning technique may involve following steps as

1. Create a training data set using sample of binaries.
2. Train classifiers to learn
3. Use the classifier to predict similarity of other binaries

Machine Learning is broadly categorized in classification and clustering. In classification, it requires training with solid truth and using and clustering representative artifacts of the classes of interest. Classification enables us to train a model on small set of known malware and trained model to find new samples in large volume of malware. So classification can be used to classify or to identify the different kinds on malware. Clustering allows us to group samples of similar behavior. In clustering different linkage metrics are used such as Average

Linkage, Centroid linkage, Complete Linkage and Median linkage. Similarly there different distance measures are used as Euclidean distance, Manhattan distance, City block distance etc.

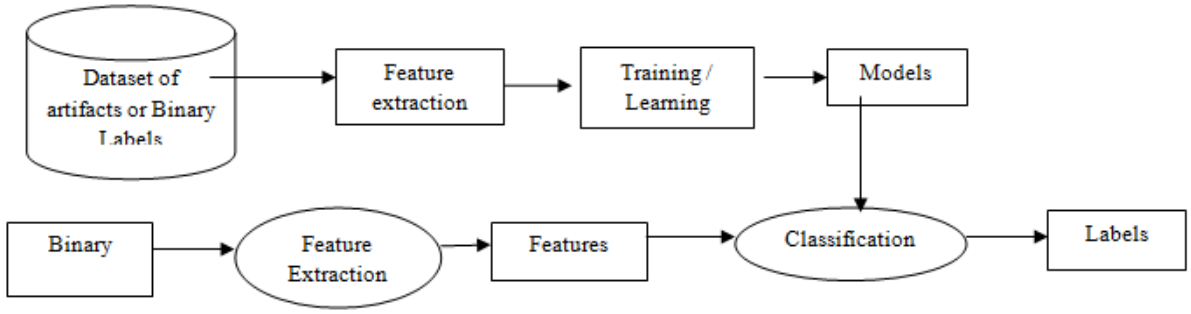

Figure 2: Diagram for Malware classification

The Fig. 2 shows the overall techniques of malware classification or clustering. **Features of Malware:** Machine learning algorithms do not directly digest raw malware, but rather first extract features that provide an abstract view of the malware. These features can be thought of as the "language" of the classifier; a way for describing malware to a given machine learning algorithm. Malware features are depending upon the class or types of malware. Static features are features extracted from the binary of the malware without executing it, i.e. through static analysis. While this can refer to the actual bits of the binary or structural information contained in the header, it more commonly refers to features extracted from the disassembled binary. Dynamic features are features that extracted by executing the malware and observing what it does, i.e. through dynamic analysis [10]. If file system is consider as a class then features can be created, modified, file size, extension etc. For registry features are created keys, modified keys and deleted keys. For IP and port class features are unique destination [7]. So as a large numbers new malware samples are arriving so that it requires an automated scheme to analysis and classify them. There are several techniques such as machine learning based, artificial Intelligence technique. Machine Learning technique for Malware classification: There are various machine learning techniques such as support vector machine, Decision tree, Naive Bayes, Association Rule based model, Clustering for detecting and classifying the malware samples. Nataraj (2011) proposed the novel method by using image processing techniques; malware can be visualized and classified which visualized the malware binaries as gray scale images. They classified malware by using K-Nearest neighbor technique with Euclidean distance [5]. They also classified malware using texture based analysis.

Depending upon based on structural information Kong, D. and Yan G(2013) was proposed Automated malware classification technique. They used function call graph as structural information. They extracted fine grained features by using function call graph for each sample. Then they have used discriminate distance metric learning to classify the malware into their respective families [1]. Tian[2008] was developed a tool for Malware classification to classify Trojans by using function length. They measure function length by numbers of bytes in the code. They have used WEKA libraries for classifying malware [2]. In 2009, Siddiqui proposed machine learning technique for detecting Worms in this paper they have used variable length instructions sequences. They used dataset of 2774 files in which 1444 files consist of worms and 1330 are benign files [5]. AMAL (2015) have introduced AMAL for large scale malware analysis, classification and clustering system. It have two subpart AutoMAL which is tool to collect low granularity behavioral artifacts which used to characterized malware usage and MaLable uses those artifacts to create representative features[7].They have used support vector machine, K-nearest neighbor and decision tree for classification and hierarchical clustering algorithms.

VI. CONCLUSION

In this paper we provide the overview in the field of malware and its analysis and classification. Malware are having similar feature. By studying it is observed that static analysis of virus is more effective than dynamic analysis. More work has done by using different techniques. Image processing techniques also used to classify malware into different its families [5]. Machine learning techniques can be used for malware classification and clustering.

REFERENCES

[1]. 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.

[2]. Tian, R., Batten, L. and Versteeg, S. (2008) Function Length as a Tool for Malware Classification. *Proceedings of the* 3*rd International Conference on Malicious and Unwanted Software*, Fairfax, 7-8 October 2008, 57-64.

[3]. Divya Bansal, Ekta Gandotra, "Malware Analysis and classification: A survey", Journal of information security , 2014.

[4]. Nataraj L, Karthikeyan, S Jacob and Manjunath, " Malware Images: Visualization and Automated Classification", Poceedings of 8th Internatinal Symposism on visualization for cyber Security, Aricle 4. 2011.

[5]. Siddiqui, M., Wang, M.C. and Lee, J. , Detecting Internet Worms Using Data Mining Techniques. *Journal of Systemics, Cybernetics and Informatics*, **6**, 48-53. 2009

[6]. Nataraj L, Yegneswaram V, Porras P and Zhang J, " A cpmrative Assement of Malware CLassifcation using binary Texture Analysis and Dynamic Analysis", Proccedings 4th ACM Workshop on Security and Artificil Intelligence, 21-30. 2011

[7]. Aziz Mohaisen, Omar Alrawi, Manar Mohaisen, "AMAL: High-fidelity, behavior based automated malware analysis and classification", Coputers & society, Elsevier. 2015

[8]. Rieck, Philipp Trinius, Carsten Willems, Thorsten Holz , Automatic Analysis of Malware Behavior using Machine Learning, Journal of Computer Security, IOS Press, http://www.iospress.nl, 2011.

[9]. https://en.wikipedia.org

[10]. C LeDoux, A Lakhotia –Malware and Machine Learning Intelligent Methods for Cyber Warfare, 2015 - Springer