Feature extraction for enhanced malware detection using genetic algorithm

Prerna Srivastava*, Mohan Raj

*Corresponding Author E-mail: prernasrivastava83@gmail.com

Abstract

The use of internet has affected almost every field today. With the increase in use of internet, the number of malwares affecting the systems has also increased to a great deal. A number of techniques have been developed by the researchers in order to detect these malwares. The Malware Detection consists of two parts, the analysis part and the detection part. Malwares analysis can be categorized into Static analysis, Dynamic analysis and Hybrid Analysis. The Detection techniques can broadly be classified into Signature based techniques and Behaviour based techniques. A brief introduction of Malware Detection techniques is addressed here. The process of Feature Extraction plays a very important role in determining the efficiency and accuracy of the Malware Detection process. It aims at determining the subset of features that helps better differentiate between the malicious and benign files. We aim to provide a Feature Extraction process based on Genetic process that can be used for Malware Detection.

Keywords: Malware; Malware Analysis; Malware Detection; Feature Extraction; Genetic Algorithm

1. Introduction

Over the past few years, the use of internet has great impact in almost every field in today’s world. Our dependence on internet has increased drastically. As the dependency on internet is increasing, the damage caused by malwares is also increasing. The widespread use of internet has given the opportunity to the malware attackers to fulfill their evil intentions. “Malwares are programs that are particularly designed to access or harm a computer without the information of the user. Malware is term signifying “malicious software.” and there are different sorts of malware including Spyware, worms etc.” [1]. As per the reports provided by the AV-TEST Institute, about 250,000 new malicious programs are registered every day. These malicious programs are then examined using the analysis tools such as Sunshine and VTEST, and then their classification is done according to their characteristics, and preserved. Then visualization programs are used to transform the results into diagrams that can be updated and current malware statistics is produced [2].

Various types of files are downloaded via internet. These files may contain different types of malwares. These malwares have the potential to add, delete or modify any existing file in the system, various information and data present on the user’s system in order to intentionally harm the system and cause damage to the information stored. These malwares intent to harm the system on which they are downloaded. They may cause undesirable consequences. The increase in the use of internet has also provided the opportunity to attackers make illegal use of the available information. In order to prevent these attacks a large number of techniques have been developed.

The term Malware consists of a combination of two terms, ‘Malicious’ and ‘Software’. Malware is short for malicious software. It aims to disrupt or gain unauthorized access to the system. There are different categories of malware that are used by the attackers (as explained subsequently). On the other hand the term cleanware refers to some software that is useful for the system, or some non-malicious software or file. It is important to clearly discriminate the malwares and cleanwares, so that none of the useful files are deleted that may result in system crash and other system failures. Various machine learning techniques have also been used to detect the presence of malware and the feature-

*Corresponding Author E-mail: prernasrivastava83@gmail.com

Fig. 1: Increase in Malware Growth
The malware detection techniques must maintain low false positive rate in order to provide better detection maintaining low training time. Malware detection system is a system that is used to determine whether a given program has some malicious intent or not. The accuracy of the classifier in correctly classifying the part depends largely on the features extracted from a particular file. Therefore careful study of these features needs to be done. The Malware Detection System comprises of two tasks – The Analysis part and the Detection part [3], which are explained in the subsequent sections.

Genetic Algorithm is a very promising field. It has been used in a number of practical problems. Genetic Algorithm is based on the concepts of survival and genetics, for evolution. This paper explores the application of Genetic Algorithm to Malware Detection. The concept of algorithm is to simulate natural evolution, specifically the principles suggested by Charles Darwin of “Survival of the Fittest”. Genetic algorithm combined with different classification technique can successfully be used for detection of Malwares [1]. Traditional Signature Based Techniques can accurately detect known Malwares but fails in case of unknown malwares. Genetic Algorithm can overcome this disadvantage by evolving new possibilities using past experiences that can help detect even unknown malwares [20]. Towards the end we will introduce method of Feature Extraction for Malware Detection, which will be based on Genetic Algorithm.

Any software that intends to harm the target system is termed as Malware. There are a number of malwares that have found to affect the systems today. As explained in [4] malwares can be classified into a number of categories based on their way of affecting the systems today. Figure 2 shows a diagrammatic representation of Malware by categories (Image source: Wikipedia).

A brief introduction to some of the most common types of malwares today is provided below:

i. **Virus**: Viruses are one of the most common type of malwares and also potentially the most harmful ones by replicating themselves. They propagate from one system to another and injects by attaching their code to some program [5].

ii. **Worms**: Worm refers to a computer program or software that replicates itself in order to infect other systems without user’s knowledge via internet. It is a program that actively transmits itself over a network in order to infect other computers and may take malicious actions [3].

iii. **Trojan Horse**: Trojan horse is a program that may appear to be legitimate but consists of hidden malicious software. Once activated, it can spy on user’s data, gain unauthorized access to data. It can perform a number of unwanted events such as adding, deleting or modifying data without user’s knowledge [6].

iv. **Spywares**: Spywares is software that is installed on the user’s system without user’s knowledge. It is considered malware since it violates user’s privacy policy. It is sometimes referred to as tracking software. It is installed on the system in order to monitor and spy on user’s activities and personal data. The personal information is collected, it can then be sent back to the attacker. This information can be then used by attacker to fulfill its evil intention [7].

v. **Adware**: Adware is a software package that consists of embedded advertisements. At times it may appear to be legitimate but may secretly be used to collect information or spy on user’s activities [3].

vi. **Ransomware**: It is a malware that gets hold of the system and demands some ransom from the victim. It restricts the user’s access to system and the data stored on it by encrypting the files, locking the system, etc.

vii. **Logic Bomb**: A Logic bomb does not publish by itself, but is installed on a system and waits until an external event is triggered such as data input, reaches to a special date, creating, deleting or even modifying a special file leading to damage of the system [2].

viii. **Backdoor**: Backdoor is a kind of malware that enters the computer system in an unauthorized way and achieves its goals without normally entering to system [4].

The rest of the paper is organized in the following way: Section II gives an overview about the Malware analysis process and its types; Section III provides an introduction to Malware Detection Techniques and its categories. Section IV explains about the Feature Extraction process. Section V presents the papers read regarding the topic. Section VI explains the methodology used and scope for future work. And finally Section VII provides the conclusion to the topic.

2. Malware Analysis

The process of analyzing a given malware sample and studying it’s properties is called, Analysis of Malware [4][19]. The Malware Analysis can be divided into two main categories: Static Analysis and Dynamic Analysis.

1) **Static Analysis**: The analysis of the software without executing it, is called Static Analysis. Static Analysis investigates the code and can detect malicious code, without running the program. It puts the investigated code in one of the available groups based on different learning methods. The low level information extracted from codes is gathered by decompiling or disassembling the codes using any of the disassembler tools such as IDA Pro and Ollydbg. Since these methods deal with real codes, they can be used in the conditions which are affected by the polymorphic malwares. One of the disadvantages of this static analysis is that, the source code of the program is usually not available which reduces the scope of usage of static analysis techniques. In the static method, binary codes are checked and viruses detected based on the binary codes.

2) **Dynamic Analysis**: In order to overcome the problems of Static Analysis, several dynamic analysis methods have been proposed. Unlike the static method that relies on malware binary codes, Dynamic Analysis is an entirely different method that analyzes the files without using their codes, but according to their...
runtime behaviour. It overcomes some of the problems of static method, but the problem with this method is that it is too slow as real time detectors on the end host and often needs the support of virtual machine technology. Analysis of the program, while it is running, is called dynamic analysis, this is also referred to as the behaviour analyzing, and it includes software running and observing its behaviour, system interaction and its effects on the host system. Dynamic analysis method runs the corrupted files in a virtual environment such as virtual machine, a simulator, sand box, etc to analyze them in a virtual environment. Several techniques have been used to analyze the programs by dynamic methods [14].

3) Hybrid Analysis: The static and dynamic analysis approaches come with their own sets of advantages and disadvantages. This has given rise to the third kind of analysis, the Hybrid Analysis, which combines the advantages of both methods [9]. It analyses the signature pattern of the input and the combines it with its behavioural patterns in order to obtain enhanced malware analysis.

3. Detection of Malware

Malware detection techniques are used to detect the malware and prevent the computer system from being infected by protecting it from potential information loss and system compromises. A number of techniques have been developed in order to detect malwares. These techniques can broadly be classified into two categories [3]: (1) Signature based techniques; (2) Behaviour based techniques [9]. The signature based techniques aims to analyze the malware sample without executing it. On the other hand, Behaviour based techniques analyze the malware samples by running it in a virtual environment.

Given below is an overview of the malware detection techniques used:

1) Signature Based Techniques: They are typically used to detect known malwares. The traditional signature based techniques maintains a list of signatures that is stored in the database. Whenever any file arrives for detection, it is checked against the signatures already stored in the database. If there is any match, an alarm is raised and the database is updated. As mentioned, a list of signature is maintained. The signature is obtained by performing the disassembly of the code. Disassembled code is first analyzed and the features are extracted. These extracted features are used to from signature for known malware families. This list needs to be constantly updated and refreshed in order to cope up with the newly arrived malwares in the market.

The main advantage of Signature-Based techniques is the accuracy it provides in detecting malwares. It can detect known instances of malware efficiently and with lesser amount of malwares required. Disadvantage of this technique is that it cannot detect unknown malwares. Whenever any new malware is introduced into the market it has to wait until the newly generated malwares harm several systems so that its signature can be formed and added to the system. Another disadvantage is that it cannot detect polymorphic or encrypted malwares.

2) Behaviour Based Detection: In this the file is made to run in a virtual environment and its behaviour is studied. Based on this analysis it is detected whether the file is malicious or benign. The main purpose of Behaviour Based Techniques is to analyze the behaviour of known or unknown malwares. It consists of two phases: training and testing phase. In training phase. Training phase consists of training the behaviour of the system is observed in the absence of malware or attack. A learning task is carried out to train the classifier with normal behaviour. In testing phase the normal phase is compared against the current behaviour. Any deviation (or anomaly) from normal behaviour is suspected as malicious activity.

The advantage of behaviour based detection is that it can detect known as well as unknown malwares. Disadvantage of this technique is that it causes slow processing of malwares and the accuracy it provides is comparatively less than that provided by signature based techniques.

Along with these, techniques are now being developed by the researchers that combine the use of data mining methods and the machine learning classifiers in order to detect malwares [10].

Data mining have been used by malware researchers for detecting the new and unknown malwares. The idea of applying the data mining and machine learning method for the detection of new malwares, was first introduced in 2001 [11], unknown malware based on their respective binary codes. The process of data mining helps in the analysis of data, by identifying meaningful patterns. The results of this analysis can be summarized to useful information which can be used for prediction. Machine learning algorithms are used for detecting patterns or relations in data, which in turn are used to develop a classifier. The general method of applying the data mining technique for malware detection is to start with generating the feature sets. The feature sets include instruction sequence, API/System call sequence, hexadecimal byte code sequence (n-gram) etc. The number of extracted features is very high therefore several text categorization techniques are applied to select only the consistent features and generate the training and test feature sets. Then the classification algorithms are applied on the trained feature set to generate and train the classifier. The performance of various classifiers is evaluated by identifying the rate of False Positive, False Negative, True Positive, True Negative, etc.

4. Feature Extraction

The Malware Detection techniques basically consist of two parts: the training part and the testing part. The entire process can be divided into two parts. The first one is the Training part and the second part is the Testing part. The Training part consists of the Feature Extraction process in which the features are extracted out of the malicious and the benign applications/files; these extracted features are used to train the classifier- whether the file send as an input is malicious or benign. Then in the testing part the trained classifier is used to check whether the inputted file is malicious or not.

Feature Extraction plays a very important role in the Malware Detection process. In this only the features that are capable of providing the best possible discriminating information are chosen. It affects the accuracy and efficiency of the algorithm to a great deal. The objective of Feature Extraction is to find that transformation, usually linear, which helps not only to eliminate irrelevant information and redundancy but also provide better separation of classes.

The large dataset of feature available can cause inefficient detection of files. Therefore Feature Extraction is performed in order to provide efficient and accurate detection of files, an important task in pattern recognition and classification problems, where only the well-chosen features provide discrimination information, and thus, can help identify the subtle changes in the machine condition. Thus feature extraction is a kind of feature selection, but also includes a transformation. The objective is typically to reduce the feature space and minimize the overall cost of measurement acquisition.

In this paper, we propose the use of the Genetic Algorithm for the feature Extraction process. The extracted features are then used to train the classifier that can be used in the Malware Detection process. The existing features are used to generate new features; these newly generated features are then evaluated using a fitness function. This method produces better detection results while preserving the low training time at the same time.

The process derives new features which are presented to the classifier, which helps improve the classification efficiency. The choice of features can greatly affect the performance of classification. Generated feature will often be refined to achieve the desired level of performance. Developing these features manually can be very time-consuming and also relies on the experience of the engineer. Generated features should be capable
of identifying the subtle or intricate relationships, within large datasets where the mapping from data to class labels is often unclear or difficult for to be identified manually.

5. Related Work

Several techniques have been developed in past in order to detect the presence of some suspicious behaviour of the file. [12] Makes use of the Genetic Programming for malware detection. A number of malwares were collected from the malware database available on the internet. Three different methods were used to perform the feature Extraction process, namely: Term frequency, Term frequency inverse document frequency and the third one Fisher Score. Two sets for balanced and unbalanced data were set up, then the results produced were compared to the ones produced by four other machine learning techniques. The machine learning techniques with which the results were compared were: Support vector machine, artificial neural network, Bayesian networks and decision trees. The results produced by GP were better than the other machine learning techniques.

Many techniques used for malware detection in Android were based on the resources required by the malware. These techniques were successful in exploring the communication between the apps and the android components, but then the inter component communication between the components were ignored. [13] Introduces the concept ICC detector for malware detection in Androids. The ICC Detector was used that exploited the Inter Component Communication (ICC) in Android to detect Malware. It comprises of two phases: The training phase and the Detection phase. In training phase various ICC based features are extracted. These extracted features are then fed as input to the classifier. The classifier is trained using these extracted features to train the classifier, the output is a detection model. The detection model is then used to differentiate between the malicious and the benign apps. [14] Introduces another malware detection technique that is used to detect whether the terminals are infected or not using the process behaviour. It makes use of two types of deep neural networks to adapt to the different characteristics of individual operational flows. The first one used is the RNN i.e. Recurrent Neural Network which is used for the feature extraction process. The second deep neural network is CNN i.e. that is Convolutional Neural Network Neural Network which is used for the feature classification step. During the training phase the API call sequences are recorded as process behaviour which is used to construct the feature extractor. The feature extractor is used to extract feature from trained RNN. These extracted features are then converted to feature images. These images are then used to train the CNN whether the input is malware or not.

Anti-malware industry mostly makes use of the stored data patterns in order to identify a malware but it fails in the case of newly generated malwares. To overcome this problem [15] uses an artificial neural network to detect malwares, which takes as input the portable executable structure from executables. They learn from the training data and are capable of identifying unknown virus patterns. Fisher score is used to identify unknown patterns and these patterns are used to train the neural network, which can then be used to discriminate infected files from legitimate ones.

The Malware Detection techniques can be broadly classified into static techniques and dynamic techniques. Static techniques make use of the signature based methodology for the detection of malwares whereas the dynamic techniques study the behaviour of the app or the file by running it in a virtual environment. On one hand where static techniques are unable to detect the unknown malware behaviours, it produces very accurate results, on the other hand dynamic techniques may overcome the disadvantage of the static techniques, but it causes very slow processing. [4] Presents a combination of the static and the dynamic techniques in order to enhance the malware detection process and help better detect the presence of malwares.

From the study of the above papers as well as others papers it can be concluded that the Feature Extraction process plays a very crucial part in the detection process. In [16], Genetic algorithm is used for the process of feature selection and feature generation. The fitness function used in the process is given by a support vector machine (svm). The error rate decreased from 11.9783 (when the svm was used only with the original features) to 0.0887. As already mentioned, feature extraction plays a vital role; therefore it should be done carefully. [17] Uses the same ideology to perform the feature extraction process using the Genetic programming.

6. Methodology Used

Various Malicious and benign files are collected from a number of sources. Then careful study of the various static and dynamic Boolean file features is done to find out the features that are helpful in determining whether the given file is useful in distinguishing between malicious and benign files or not. When it comes to Feature Extraction using Genetic Algorithm, it can be done using two methods: (1) first, in which classifier is part of the fitness function. Though this method may produce better results than the other method but the disadvantage of this approach is that a lot of time is wasted in training the classifier since it is done alongside the fitness function [16]. (2) In the second method, first the features are extracted or evolved using the original features, and then these extracted features are evaluated using an appropriate fitness function and only once the Genetic Algorithm is completed, the resulting features are used to train a classifier [4].

The extracted features help detection techniques to analyse the features of various malicious and benign files. As described in [17], the file features, form the chromosome population. The chromosome is composed of complex genes. Each gene consists of three components (As depicted in figure 3): the feature vector (F), a Boolean operator (B) and a priority (P).

![Fig. 3: Chromosome Structure](image-url)

The Boolean operator used are mainly AND, XOR and NOT. The operators AND and XOR are used for creating new chromosomes. The NOT operator is used for mutation. For each generation two chromosomes are randomly selected which form the parent chromosomes. The cross-over operators are applied on the two parent chromosome to form new chromosomes, called offsprings. Mutation may be applied on a chromosome by randomly changing the components of genes to produce new offspring. Process is repeated till a given termination condition is reached. Then at the end, when the termination condition is reached, extracted features are used to train the classifier. This trained classifier is now capable of differentiating between the malicious and benign files. The accuracy of the extracted feature depends greatly on the fitness function chosen for the purpose. Therefore appropriate selection of the fitness function is important.
In the earlier versions of genetic algorithm that was used for feature extraction, the results were affected by the ‘bloating problem’ [17]. This paper addresses the problem of bloating that occurs in the chromosome structure. The bloating problem causes variation in the size of the chromosome structure. This sufficiently causes the decrease in the efficiency of the algorithm therefore it needs to be prevented. There are a number of solutions that are provided to the bloating problem [18]. Some of them are: (1) limiting the size and also the depth of the individuals during the population initialization phase. (2) By adding some kind of tree size penalty which can be used for the fitness function evaluation (3) The offspring or the children generated in each generation are included in the next generation only if the value of the resulting offspring is either greater than or equal to the value of the parents. In this paper, it is done by attaching an extra field which we have referred over here as ‘weight’ (represented as W in the chromosome structure depicted in figure 4), to the chromosome that indicates the frequency or the number of malicious or benign files in which the feature is set to true/false. The calculation of the fitness function is done based on this field.

Fig. 4: Additional field in chromosome structure

Summarising what is said above; the paper makes use of the Genetic Algorithm for the Feature Extraction process. Various malicious files are collected via different sources available on internet. The extracted features are then used to evolve new features which are evaluated using a fitness function. The process is repeated until a given termination condition is satisfied. At the end the evolved features are used to train the classifier that can be used in the Malware Detection process (Figure 5). These features create the possibility of detecting unknown malwares that may occur in future. Thus, this method produces better detection results while preserving the low training time at the same time.

7. Conclusions

This paper provides a brief introduction to types of malware, Malware analysis methods and techniques of Malware Detection. Various techniques are used for the Detection of Malware, which can broadly be classified into Signature Based Techniques and Behaviour Based techniques. Recent developments have also been made that applies the use of data mining methods and machine learning based technology for Malware Detection. Feature Extraction plays a very important role in the Detection process. It greatly affects the accuracy and efficiency of the process. Therefore it needs to be done carefully. In the use of Genetic Algorithm for Feature Extraction process, the existing features are used to generate new features; these newly generated features are then evaluated using a fitness function. The extracted features are used to train the classifier that can be used in the Malware Detection process. These features create the possibility of detecting unknown malwares that may occur in future. The use of Genetic Algorithm in feature extraction produces better detection results while preserving the low training time at the same time.

References


