Data Mining for Malware Detection

Surya Alla

I. ABSTRACT

Data mining techniques have been concentrated for malware detection in the recent decade. The battle between security analyzers and malware scholars is everlasting as innovation grows. The proposed methodologies are not adequate while evolutionary and complex nature of malware is changing quickly and therefore turn out to be harder to recognize. This paper presents a systematic and detailed survey of the malware detection mechanisms using data mining techniques. In addition, it classifies the malware detection approaches in two main categories including signature-based methods and behavior-based detection.

II. INTRODUCTION

In the recent years, the application of malware detection mechanisms utilizes through data mining techniques through have increased using machine learning to recognize malicious files. Machine learning methods can take in hidden examples from a given preparing set which includes both malware and benign examples. These basic examples can separate malware from benevolent code. Malware is a standout most thoughtful intimidations for distributed systems and the Internet. The battle between security analyzers and malware scholars is everlasting as innovation grows. Malware is a program that makes your framework accomplish something that an assailant needs it to do. The most generally utilized malware detection develops a straightforward example coordinating way to deal with identify vindictive code. Typically, malware designers don't compose new code without any preparation, yet redesign the old code with new components or muddling strategies. With a large number of malware cases seeming each day, proficiently preparing countless specimens which display comparable conduct, has turned out to be progressively essential.

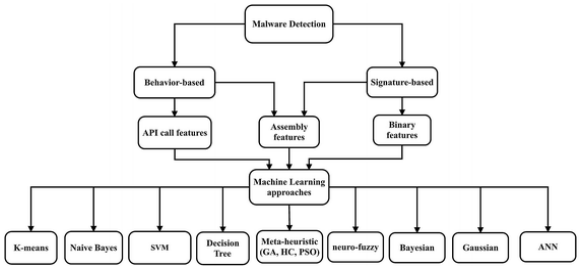
Up to now, malware analysis has the high growing impact in the procedure of deciding the reason and the usefulness the conduct of a given suspicious application. Such a procedure is an important essential with a specific end goal to create effective and powerful identification furthermore characterization techniques; malware analysis is partitioned into two primary classifications that include dynamic and static methods.

This review classifies the malware detection approaches in two main fields: signature-based and behavior-based.

III. MALWARE DETECTION APPROACHES

As a result of the developing malware in the innovation, the information of obscure malware protection is a fundamental subject in the malware recognition as per the machine learning strategies. The machine learning strategies are divided into supervised and unsupervised classes. Malware detection approaches are divided into two main categories that include behavior-based and signature-based methods.

In Fig we illustrate a malware detection taxonomy based on machine learning approaches. According to this figure, the API calls features, assembly features, and binary features are existing approaches for malware detection method. These features use machine learning methods for predicting and detecting malicious files.

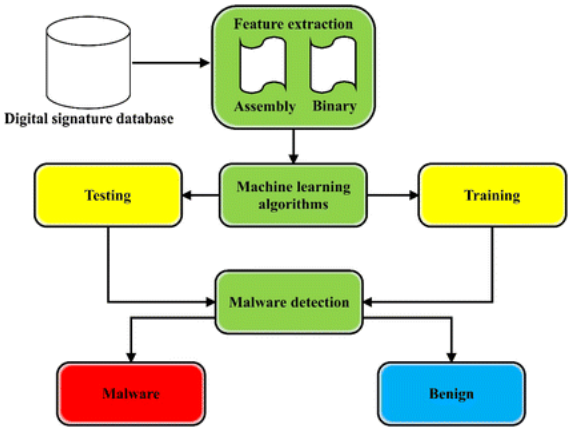


- **Signature Based Detection**

Recently, signature-based detection is the most generally utilized procedure in antivirus programming highlighting exact correlation. Malware recognition has essentially centered on performing static investigations to review the code-structure mark of infections, instead of element behavioral methods. The signature-based system finds interruptions utilizing a predefined list of known assaults. Despite the fact that this arrangement has the ability to identify malware in the versatile application, it requires steady overhauling of the predefined signature database. Moreover, it is less effective in identifying noxious exercises utilizing the signature-based technique because of the quickly changing nature of portable malware. Signature-based strategies depend in light of exceptional crude byte examples or standard articulations, known as marks, made to coordinate the noxious document. For example, static highlights of a record are utilized to decide if it is a malware. The main advantage of signature-based techniques is their thoroughness since they follow all conceivable execution ways of a given document.

In inside of the malware structure, existing malicious objects have characteristics that can be used to generate a unique digital signature. The anti-malware provider utilizes the meta-heuristic algorithms that can scan efficiently the malicious object to control its signature. After identifying the malicious object, the detected signature is added to the existing database as the recognized malware. The database sources include huge number of the various signatures that classify malicious objects. In the signature-based malware detection, there are some various qualities including fast identification, easy to run, and broadly accessible.

Since the digital signature plans are gotten from known malware, these plans are likewise generally known. Subsequently they can be effectively evaded by programmers utilizing straightforward confusion procedures. Hence malware code can be modified, and signature-based identification can be sidestepped. Since anti-malware providers are built on the premise of known malware, they can't to distinguish obscure malware, or even variations of known malware. In this way, without exact digital signature, they can't adequately distinguish polymorphic malware. Along these lines, signature-based recognition does not give zero-day insurance. Besides, since a signature-based indicator utilizes an isolate signature for each malware variation, the database of signatures develops at an exponential rate. The signature-based malware detection has two main methods for applying malware detection approach in machine learning methods including assembly features and binary features. The below figure is a depiction of signature-based method



Advantages:

- Easy to run
- Fast Identification
- Broadly Accessible
- Finding comprehensive malware information

Disadvantages:

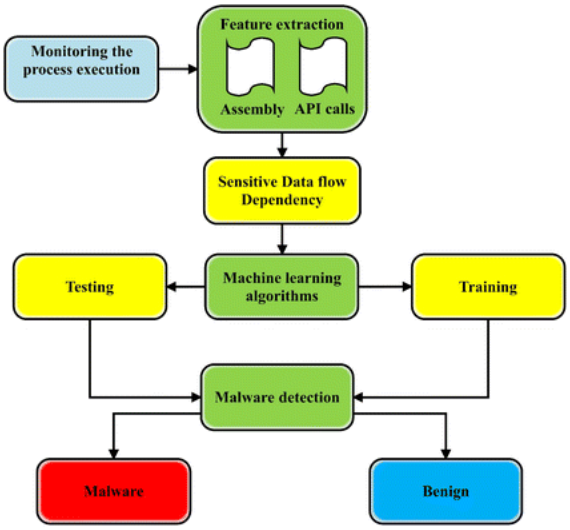
- Failing to detect the polymorphic malwares
- Replicating information in the huge database

- **Behavior based Detection**

Behavior-based methodologies require execution of a given example in a sandboxed situation and run-time exercises are checked and logged. Dynamic investigation systems utilize both

virtualization and imitating conditions to execute a malware and to remove its practices. The primary advantage of the behavior-based approach is that gives a superior comprehension of how malware is produced and implemented.

In the behavior-based malware approach, the suspicious objects are assessed based on their activities that they cannot execute in system. Efforts to achieve activities that are clearly irregular or unofficial would specify the suspicious object is malicious, or at least apprehensive. A malicious behavior is known using a dynamic analysis that evaluates malicious intent by the object's code and structure. In the behavior-based detection, the API calls and assembly features are two main methods for applying machine learning algorithms.



Advantages:

- Detecting unconceived types of malware attacks
- Data-flow dependency detector
- Detecting the polymorphic malwares

Disadvantages:

- Storage complexity for behavioral patterns
- Time complexity

IV. Review of the approaches

- **Signature Based Detection**

Cui et al. [2] illustrated a novel recognition framework in light of cloud environment and packet examination. The framework identifies the malicious mobile malware behavior through their bundles with the utilization of information mining strategies. This approach totally keeps away from the deformities of customary techniques. The framework is administration arranged and can be sent by portable administrators to send cautions to clients who have malware on their gadgets. To enhance framework execution, another bunching technique called withdrawal grouping was made. This technique utilizes earlier learning to lessen dataset measure. In addition, a multi-module location plan was acquainted with improve framework precision. The aftereffects of this plan are created by incorporating the location consequences of a few calculations, including Naive Bayes and Decision Tree.

Hellal and Ben Romdhane [3] displayed another diagram mining technique to recognize variations of malware utilizing static examination while covering the current defects. Also, they proposed a novel calculation, called minimal contrast frequent sub-graph miner method (MCFSM), for separating negligible discriminative and generally utilized malevolent behavioral designs which can distinguish definitely a whole group of vindictive projects, conversely to another arrangement of benevolent projects. The proposed technique demonstrates high recognition rates and low false positive rates and creates a predetermined number of malware marks.

Based on the research papers, the main advantage of signature-based detection approaches is using pattern detection that decreases the system overhead and execution time for malware prediction. The main disadvantage of the signature-based detection approaches is omitting feature selection.

- **Behavior based Detection**

Yuan et al. [4] presented a deep learning method to connect the components from the static investigation with elements from the dynamic investigation of Android applications. In addition, they actualized an Android malware detection engine based on the deep-learning method (DroidDetector) that can consequently distinguish whether a file has a malicious behavior or not. With a large number of Android applications, they tested DroidDetector and play out an in-depth examination of the elements that deep learning basically adventures to portray malware completely. The outcomes appear that deep learning is appropriate for characterizing Android malware and particularly compelling with the accessibility of additional preparation information. DroidDetector can accomplish 96.76% detection accuracy, which traditional machine learning methods.

Boukhtouta et al. [5] presented the issue of fingerprinting perniciousness of activity with the end goal of recognition and arrangement. This research pointed first at fingerprinting perniciousness by utilizing two approaches: Deep Packet Inspection (DPI) and IP bundle headers arrangement. To this end, we consider malignant activity created from element malware examination as movement perniciousness ground truth. In light of this supposition, they exhibited how these two methodologies are utilized to recognize what's more, attribute maliciousness to the various threat. In this work, we concentrate the positive and negative angles for Deep Packet Review and IP bundle headers order. They assessed every approach in view of its recognition and attribution precision and additionally their level of multifaceted nature. The results of both methodologies have demonstrated promising outcomes as far as discovery; they are great possibility to constitute a collaboration to expand or prove recognition frameworks as far as runtime speed and grouping exactness.

Based on the research papers, the main advantage of behavior-based detection approaches is detecting all of the suspicious files according to their calls' behavior that increases the accuracy of malware prediction. The main disadvantage of the behavior-based detection approaches is the runtime overhead.

V. Summary

In conclusion, there are 2 type of malware detection approaches signature based the easier to run and faster identification approach and behavior based which detect more unconceived types but takes up more space and time, both of which are carried out using various machine learning algorithms like decision tree etc. As showed above, both the approaches have their own advantages and disadvantages making them viable in different scenarios. Based on some of the research papers in this topic, Signature Based Detection is better suited for windows-based applications and behavior-based detection is better suited for smartphones.

VI. References

1. https://hcis-journal.springeropen.com/articles/10.1186/s13673-018-0125-x#Sec5
2. Cui B, Jin H, Carullo G, Liu Z (2015) Service-oriented mobile malware detection system based on mining strategies. Pervasive Mob Comput 24:101–116.
3. Hellal A, Romdhane LB (2016) Minimal contrast frequent pattern mining for malware detection. Comput Secur 62:19–32.
4. Yuan Z, Lu Y, Xue Y (2016) Droiddetector: android malware characterization and detection using deep learning. Tsinghua Sci Technol 21:114–123.
5. Boukhtouta A, Mokhov SA, Lakhdari N-E, Debbabi M, Paquet J (2016) Network malware classification comparison using DPI and flow packet headers. J Comput Virol Hacking Tech 12:69–100.