International Journal of e-Collaboration
Volume 18 • Issue 1

Malware Analysis Using Classification and Clustering Algorithms

Balaji K. M., Vellore Institute of Technology, Chennai, India
Subbulakshmi T., Vellore Institute of Technology, Chennai, India

ABSTRACT

Malware analysis and detection are important tasks to be accomplished as malware is getting more and more arduous at every instance. The threats and problems posed by the public around the globe are also rapidly increasing. Detection of zero-day attacks and polymorphic viruses is also a challenging task to be done. The increasing threats and problems lead to the need for detection techniques which lead to the well-known and the most common approach called machine learning. The purpose of this survey is to formulate the most effective feature extraction and classification ways that sums up the most effective methods (which includes algorithms) with maximum accuracy and also to effectively understand the clustering properties of the malware datasets by considering appropriate algorithms. This work also provides an overview on information about malwares used. The experimental results of the proposed model clearly showed that the KNN classifier as the most accurate with 0.962355 accuracy.

KEYWORDS

Dynamic and Integrated Malware Analysis, K-Means Clustering, K-Nearest Neighbour Classifier, Machine Learning, Malicious, Static

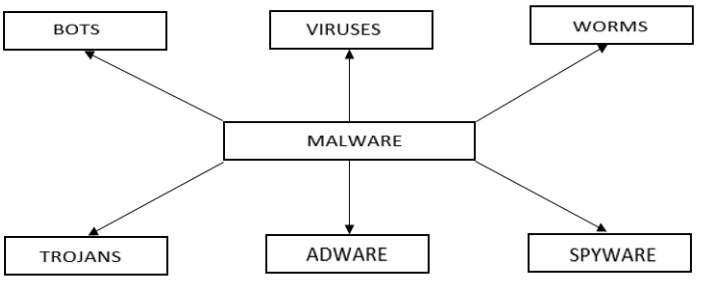
INTRODUCTION

The threat that malicious software is causing to the digital world is growing rapidly. As per AV-TEST, the aggregate number of new malware tests is expected to outperform 700 million by 2020 (AV-TEST, 2020). It is nearly impossible to control such massive amount of malwares. Therefore, networking and security researchers are using malware identification and detection systems to detect the malwares which initially includes two stages that is detection and analysis. This can be achieved through static or dynamic or integrated approach. The main goal of malware analysis is to record and capture the properties which can be additionally used to improve the security measures and make evasion of malware as difficult as possible. Figure 1 shows the different classification of malwares and these malwares can be present in any form or category such as a script, a segment of code or any other binary. The purpose of malware is to get the control of the system, derange the services of computer systems, take back the available functions, rob the restricted information and damage the sources.

Illegal applications sometimes act as protective cover for the malwares. Trying to gain access to this illegalised software from many websites may download the malware itself. In general, this case

DOI: 10.4018/IJeC.290290

Figure 1. Classification of malware



is possible and found in cracked/pirated software. These malicious software are not only operatable, executable source codes but can act as supportive downloaders for malicious files like portable document formats (PDF) or other links. As per VirusTotal, 47.80% of malicious files are executables (More than 100M files with original information; more than 16M portable executables from distinct URLs; more than 20M files with rich telemetry data; more than 700,000 emails for rich contextual information). So, the intention here is to dissect these executables. Numerous malwares are available, and they can be categorized into Trojan pony, Virus, Worm, Adware, and Backdoor. Few of them cannot be arranged into a particular group, because malwares have various attributes which helps them to coordinate in various classifications and at some point, they are referred as generalized malicious files. Malware files are dissected on the methods of dynamic and static techniques.

Figure 2 shows us the statistical records gathered from the cybercrime magazines (Morgan, 2019) on the ransomware attacks on business. It is estimated that the total damage costs shall exceed 20 Billion by 2021 and expected to attack a business every 11 seconds by the end of 2021.

The three basic analysis methods to analyse malware are as follows.

Static Analysis

Static analysis is a method in which the executable documents and files are tested for malware without executing it in an environment that is dynamically controlled. Executable files have numerous statistical features such as segments and memory minimization. The PE file format is a library in python which removes static highlights even in the presence of executable records.

Figure 2. Records obtained from cybercrime magazine