Chapter 13
Comparative Analysis of Feature Selection Methods for Detection of Android Malware

Meghna Dhalaria
Jaypee University of Information Technology, India

Ekta Gandotra
Jaypee University of Information Technology, India

Deepak Gupta
https://orcid.org/0000-0002-1547-196X
Jaypee University of Information Technology, India

ABSTRACT

Over the past few years, Android has been found to be the most prevalent operating system. The increase in the adoption of Android by users has led to many security issues. The amount of malware targeting Android has significantly increased. Due to the increase in the amount of malware, their detection and classification have become a major issues. Currently, the detection techniques comprise static and dynamic malware analysis. This chapter presents a comparative study of various feature selection methods through machine learning classifiers for Android malware classification. The study examines the features acquired through static malware analysis (such as command strings, permissions, intents, and API calls), and various feature selection techniques are employed to find suitable features for classifying malware to carry out the comparative analysis. The experimental results illustrate that the gain ratio feature selection approach selects relevant features for the classification of Android malware and provides an accuracy of 97.74%.

DOI: 10.4018/978-1-6684-6275-1.ch013

INTRODUCTION

In the current era, there is an increase in the usage of smartphones in our day-to-day lives. A Lot of users use these smartphones due to the various functionalities they are providing like emailing, gaming, watching videos, banking etc. There are various operating systems (OS) available in the market such as Windows, iOS, Android, BlackBerry etc. Among these, Android is found to be the most prevalent one. Android is a mobile OS based on Linux kernel specifically built for touchscreen devices such as tablets and smartphones etc. It is established in 2005 by Google (Android Developers, 2011). Unfortunately, the renown of these devices has resulted in an increasing level of cybercrimes. Malware is considered the major issue with respect to security threats. It refers to distinct forms of intrusive software such as worms, rootkits, backdoors and trojan horses etc. They perform malicious activities like extracting unauthorized personal information, destruct the information or gain access to a device etc. In various attack scenarios, an attacker can exploit Android's vulnerabilities and compromise a user. In one scenario, a Trojan app might provide cool HD wallpapers in the foreground, while secretly collecting private information from users' phones, such as contacts. A wallpapers app will require INTERNET permission in order to download wallpapers. Unsuspecting users may not check permission requests and accidentally grant READ_CONTACTS permission. The attacker can use this data for monetary gain and/or propagation of malware. Another scenario involves an attacker draining a victim's smartphone's battery life by overusing resource-consuming services like radio, GPS, etc. to kill their smartphone.

According to the worldwide Statista report, Android is turned out to be a leading mobile OS with 74.13% of market share by December 2019 (Statista: Mobile operating systems market share, 2020). The increase in the use of Android resulted in the increase of malware. In 2020, McAfee stated that there is 121 million total and 49 million new malware are discovered (McAfee Lab: Threat Predictions Report, 2020).

The increasing use of Android applications (apps) lured attackers to build complex and sophisticated malware which is difficult to analyze. The earlier signature-based approach was extensively used for identifying malware. This method extracts patterns from the fishy files matching with the malware signature database for malware detection (Kouliaridis et al., 2020; Aslan and Samet, 2020). The major constraint of this method is that it cannot detect unknown malware (i.e. zero-day). Machine learning (ML) techniques are now being used by researchers to detect Android malware. These methods allow computers to think and make predictions. Static and dynamic malware analysis features can be used by machine learning approaches to classify apps (Memon and Anwar, 2015, Dhalaria and Gandotra, 2021a; Dhalaria and Gandotra, 2020a; Dhalaria and Gandotra, 2022; Dhalaria and Gandotra, 2020b; Dhalaria and Gandotra, 2021b).

Static malware analysis investigates the app by investigating its source code. The major constraint of this approach is that it is unable to examine the obfuscation code (Barrera et al., 2010; Singla et al., 2015). But this limitation is overwhelmed by dynamic malware analysis which monitors the activities of an app by executing it in a sandbox (virtual environment). The major constraint of this approach is that it needs more time and resources and cannot explore all execution paths (Gandotra et al., 2014). A more number of features degrade the accuracy of the classification models and may increase the complexity of the model. So to overcome this problem, different feature selection techniques are used. These techniques help in selecting appropriate features for the purpose of classification. Moreover, it also reduces the model building time.