Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Discovering optimal features using static analysis and a genetic search based method for Android malware detection*

Ahmad FIRDAUS†‡1,2, Nor Badrul ANUAR†‡1, Ahmad KARIM[3], Mohd Faizal Ab RAZAK[1,2]

[1]*Department of Computer System and Technology, University of Malaya, Kuala Lumpur 50603, Malaysia*
[2]*Faculty of Computer System & Software Engineering, University Malaysia Pahang, Gambang 26300, Malaysia*
[3]*Department of Information Technology, Bahauddin Zakariya University, Multan 60000, Pakistan*
†E-mail: ahmadfirdaus@um.edu.my; badrul@um.edu.my

Received Aug. 22, 2016; Revision accepted Mar. 15, 2017; Crosschecked June 8, 2018

Abstract: Mobile device manufacturers are rapidly producing miscellaneous Android versions worldwide. Simultaneously, cyber criminals are executing malicious actions, such as tracking user activities, stealing personal data, and committing bank fraud. These criminals gain numerous benefits as too many people use Android for their daily routines, including important communications. With this in mind, security practitioners have conducted static and dynamic analyses to identify malware. This study used static analysis because of its overall code coverage, low resource consumption, and rapid processing. However, static analysis requires a minimum number of features to efficiently classify malware. Therefore, we used genetic search (GS), which is a search based on a genetic algorithm (GA), to select the features among 106 strings. To evaluate the best features determined by GS, we used five machine learning classifiers, namely, Naïve Bayes (NB), functional trees (FT), J48, random forest (RF), and multilayer perceptron (MLP). Among these classifiers, FT gave the highest accuracy (95%) and true positive rate (TPR) (96.7%) with the use of only six features.

Key words: Genetic algorithm; Static analysis; Android; Malware; Machine learning
https://doi.org/10.1631/FITEE.1601491 **CLC number:** TP309.5

1 Introduction

Numerous people use mobile devices frequently in their daily activities to accomplish imperative tasks, such as health management, synchronous data transfer, family communications, money transactions, and sensor activities (La Delfa et al., 2016). Among all mobile device operating systems (OSs), Android dominates the smartphone market with a share worth of 366 billion dollars (Thomas, 2015). Furthermore, Android devices are available in a wide range of prices from as low as 50 dollars (eBay, 2016). Given that Android appliances are ubiquitous, Android malware has been rapidly growing in scale. Unscrupulous programmers design malware computer programs to perform diverse malicious actions without users' consent. These actions infiltrate user devices to gain data (e.g., photos, bank activities, messages, phone contacts, and map locations) (Karim et al., 2014). The Sophos mobile security website discovered 610 389 new malware samples outside of the Google Play market in the first six months of 2015 (Komili, 2015). Also, security analysts have discovered hidden malware in 104 applications in the Google Play store; these applications have been downloaded over 3.2 million times in user devices (Russon, 2016). Thus, detecting malware is imperative, as well as countering and reducing its threats and

‡ Corresponding author

* Project supported by the Ministry of Science, Technology and Innovation of Malaysia, under the Grant eScienceFund (No. 01-01-03-SF0914)

ORCID: Ahmad FIRDAUS, http://orcid.org/0000-0002-7116-2643