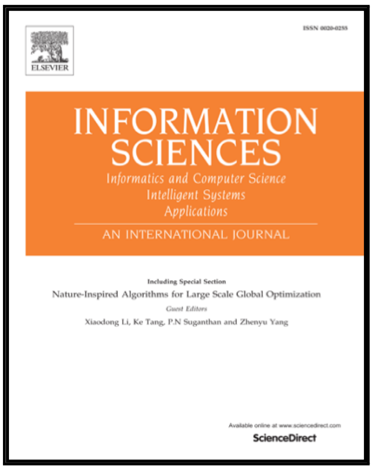

ISSN 0020-0255

ELSEVIER

INFORMATION
SCIENCES
Informatics and Computer Science
Intelligent Systems
Applications

AN INTERNATIONAL JOURNAL

Including Special Section
Nature-Inspired Algorithms for Large Scale Global Optimization
Guest Editors
Xiaodong Li, Ke Tang, P.N Suganthan and Zhenyu Yang

Available online at www.sciencedirect.com
ScienceDirect

Machine Learning Based Mobile Malware Detection Using Highly Imbalanced Network Traffic

Zhenxiang Chen, Qiben Yan, Hongbo Han, Shanshan Wang, Lizhi Peng, Lin Wang, Bo Yang

PII:	S0020-0255(17)30707-7
DOI:	10.1016/j.ins.2017.04.044
Reference:	INS 12865

To appear in:	*Information Sciences*

Received date:	15 March 2016
Revised date:	8 April 2017
Accepted date:	28 April 2017

Please cite this article as: Zhenxiang Chen, Qiben Yan, Hongbo Han, Shanshan Wang, Lizhi Peng, Lin Wang, Bo Yang, Machine Learning Based Mobile Malware Detection Using Highly Imbalanced Network Traffic, *Information Sciences* (2017), doi: 10.1016/j.ins.2017.04.044

Available online at www.sciencedirect.com

ScienceDirect

Information Sciences 00 (2017) 1–22

INS

www.elsevier.com/locate/procedia

Machine Learning Based Mobile Malware Detection Using Highly Imbalanced Network Traffic

Zhenxiang Chen[a,b], Qiben Yan[c,*], Hongbo Han[a,b], Shanshan Wang[a,b], Lizhi Peng[a,b], Lin Wang[a,b], Bo Yang[b,*]

[a] *School of information science and engineering University of Jinan, Jinan 250022, China*
[b] *Shandong Provincial Key Laboratory of Network Based Intelligent Computing, Jinan 250022, China*
[c] *University of Nebraska-Lincoln, Lincoln, NE 68588, USA*

Abstract

In recent years, the number and variety of malicious mobile apps have increased drastically, especially on Android platform, which brings insurmountable challenges for malicious app detection. Researchers endeavor to discover the traces of malicious apps using network traffic analysis. In this study, we combine network traffic analysis with machine learning methods to identify malicious network behavior, and eventually to detect malicious apps. However, most network traffic generated by malicious apps is benign, while only a small portion of traffic is malicious, leading to an imbalanced data problem when the traffic model skews towards modeling the benign traffic. To address this problem, we introduce imbalanced classification methods, including the synthetic minority oversampling technique (SMOTE) + support vector machine (SVM), SVM cost-sensitive (SVMCS), and C4.5 cost-sensitive (C4.5CS) methods. However, when the imbalance rate reaches a certain threshold, the performance of common imbalanced classification algorithms degrades significantly. To avoid performance degradation, we propose to use the imbalanced data gravitation-based classification (IDGC) algorithm to classify imbalanced data. Moreover, we develop a simplex imbalanced data gravitation classification (S-IDGC) model to further reduce the time costs of IDGC without sacrificing the classification performance. In addition, we propose a machine learning based comparative benchmark prototype system, which provides users with substantial autonomy, such as multiple choices of the desired classifiers or traffic features. Using this prototype system, users can compare the detection performance of different classification algorithms on the same data set, as well as the performance of a specific classification algorithm on multiple data sets.

Keywords: Network traffic; Malicious apps; Imbalanced data; Malware detection; Machine learning

1. Introduction

With the rapid growth of mobile networks, the number of mobile devices surges to a record high. However, mobile users still face serious threats from ransomware, spyware, malicious apps, and financial malware. Modern mobile devices are no longer confined to traditional communication services, and the popularity of apps related to e-commerce, personal payments, and social communication on mobile devices is continuously increasing. Security issues in mobile devices, particularly app security on Android platforms,

[*] Corresponding author. E-mail addresses: yan@unl.edu (Q. Yan), yangbo@ujn.edu.cn (B. Yang)