ANALYSIS OF CLASSIFICATION ALGORITHMS FOR ANDROID MALWARE BEHAVIOUR DETECTION

MUHAMMAD SAUFI BIN SIRON

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS TESIS*

JUDUL : ANALYSIS OF CLASSIFICATION ALGORITHMS FOR ANDROID MALWARE BEHAVIOUR DETECTION

SESI PENGAJIAN : 2013 / 2014

Saya MUHAMMAD SAUFI BIN SIRON

mengaku membenarkan tesis Projek Sarjana Muda ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. ** Sila tandakan (/)

_____ SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
_____ TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
_____ TIDAK TERHAD	

(TANDATANGAN PENULIS)	(TANDATANGAN PENYELIA)
Alamat tetap: No 26 jalan Selaseh 3,	PN. SYARULNAZIAH BINTI ANAWAR
Taman Selaseh Fasa 1,	
68100 Batu Caves, Selangor	
Tarikh	Tarikh:

CATATAN: * Tesis dimaksudkan sebagai Laporan Projek Sarjana Muda (PSM).
** Jika tesis ini SULIT atau atau TERHAD, sila lampirkan Surat daripada pihak berkuasa.

ANALYSIS OF CLASSIFICATION ALGORITHMS FOR ANDROID MALWARE BEHAVIOUR DETECTION

MUHAMMAD SAUFI BIN SIRON

This report is submitted in partial fulfillment of the requirements for the Bachelor of Computer Science (Computer Networking)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA
2014

DECLARATION

I hereby declare that this project report entitled
ANALYSIS OF CLASSIFICATION ALGORITHMS FOR ANDROID MALWARE BEHAVIOUR DETECTION

is written by me and is my own effort and that no part has been plagiarized without citations.

STUDENT: _____ Date :_____
(MUHAMMAD SAUFI BIN SIRON)

SUPERVISOR: _____Date :_____
(PN. SYARULNAZIAH BINTI ANAWAR)

DEDICATION

Dear Parent

Thank you for your sacrifice and love.

Dear Teachers and Supervisors

Thank you for all the knowledge and guidance.

Dear Friends

Thank you for all the knowledge, guide, encouragement and love.

Acknowledgement

First of all, I would like to express my very great appreciation to my supervisor, Pn. Syarulnaziah Binti Anwar and Mr. Mohd Zaki Mas'ud for guide and give useful critiques for this Analysis Android Malware Detection through Machine Learning project. I would like to thanks to her because give advice and assistance in keeping my progress on schedule. I would also like to thanks to my friends Zhong wei, Nabila and all fellows friend from BITC student for helping me to complete my project. Finally, I wish to thank to my parents for their support and encouragement throughout my study

Abstract

Today more malware has been created for new target which is smart phone with android platform. So with this project it will help to other researcher to detect behavior of malware using classifier technique. There are several problems to be solved through this project which is, there is no formal procedure to perform data collection using android platform that cause a problem to other researcher to collect data for analysis in future, and there are no comparative study on malware behavior through android platform. Then, the objective of this project to answer the problem which is to propose procedure to perform data collection on android platform, to determine the malicious behavior of malware in android and last to find the best classification technique for android malware detection. This project contribution to for those want to know about the malware process on android application based on syscall process, beside that this research also help other researcher to guide them how to implement data collection on android platform and to choose the best classification to used for android malware data. Then from the result and finding shows the best technique and algorithm used to get high accuracy this android data

Abstrak

Hari ini, lebih virus telah dicipta untuk sasaran baru iaitu telefon pintar dengan belandaskan android. Dengan itu projek ini ia akan membantu penyelidik lain untuk mengesan tingkah laku malware menggunakan teknik pengelas. Terdapat beberapa masalah yang perlu diselesaikan melalui projek ini, tidak ada prosedur yang formal untuk melaksanakan pengumpulan data menggunakan platform android yang menyebabkan masalah kepada penyelidik lain untuk mengumpul data untuk analisis di masa depan, dan tidak ada kajian perbandingan ke atas tingkah laku malware melalui platform android. Kemudian, objektif projek ini untuk menjawab segalam masalan dan persoalan yand ada dimana untuk mencadangkan prosedur untuk melaksanakan pengumpulan data pada platform android, untuk menentukan tingkah laku ataupun sifat yang ada malware pada android dan terakhir untuk mencari teknik pengelasan terbaik untuk mengesan malware pada android. Sumbangan projek kepada orang-orang ingin tahu tentang proses malware pada android aplikasi berasaskan proses syscall, selain itu kajian ini juga membantu penyelidik lain untuk menunjukkan mereka bagaimana cara untuk melaksanakan pengumpulan data pada platform android dan memilih klasifikasi yang terbaik untuk digunakan untuk android data malware. Kemudian dari hasil dan keputusan menunjukkan teknik terbaik dan algoritma yang digunakan untuk mendapatkan ketepatan yang tinggi pada data android ini

.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Introduction

1.1 Introduction

The first chapter of this project will discuss about the project background, problem statement, objective, scopes, project significances, expected output from this work and Report Organization. This chapter will explain more detailed about the project.

1.2 Project Background

This project focus on classifier algorithm for android malware, classification is an ordered set of related categories used to know the similarities of data between same classes. This technique to descriptors and allows survey response to be put into meaningful categories in order to have the useful data. On this project it has 2 classes which are malware and clean. All type of data can be categorized easily because of this technique.

The reason this project is significant are because everybody can put their own application to allow other user to download, and then it will cause the limited control on malware application has been distributed. There are too many tool for repacking existing application with malicious code inside that, because of that, too difficult to detect malware application using know day tool. Because of that classification can used in future to test with new application for classifier whether that application malware or not. Beside that this project is significant because, to help and guide other researcher on several issues such as technique to chose for android malware data and the ways to perform data collection

In this project, it will focus on classification technique to evaluate the performance of the technique, and our target is to investigate the performance of classification techniques with different algorithm. Classifier is a supervised learning and requires training data are accompanied by labeling the class, and new data is classified based on

the training set. This project wants to evaluate the best performance of classification that can be used for real world environment and experiment.

1.3 Problem Statement

In this part it will discuss about the current problem need to solve by this project. There are several classifier techniques through machine learning, but there are not has been prove which technique will produce the best result and performance for android malware behavior detection.

Table 1.1: Problem Statement

No	Project Problem
PP1	There is no formal procedure to perform malware data collection on Android platform that will cause problem for researcher to perform data collection. Because of different environment.
PP2	There is no comparative study on malware behavior detection on Android platform using classification technique. This will cause problem for other researcher to use which classify on android malware data.

.

1.4 Project Question

Table 1.2: Project Question

PP	PQ	Project Question (PQ)
PP1	PQ1	How to collect data for classifier on android platform?
	PQ2	Which classifier will produce the accurate result to detect malware?
	PQ3	Which classifier needs to choose by investigator to detect malware behavior?
PP2	PQ4	Which classifier will produce the best performance on malware behavior detection?

1.5 Project Objective

Table 1.3: Project Objective

PP	PQ	PO	Project Objective (PO)
PP1	PQ1	PO1	To propose formal procedure to perform data collection on android platform
	PO2 PQ3	PO2	To determine the malicious behavior of malware in android platform
PP2	PQ1	PO3	To determine the best classifacation technique for malware behavior detection.

PO1 to produce the guideline for investigator to perform data collection on android operating system or any device that use android operating system.

PO2 the objective is to determine which system call or behavior that creates by Smartphone process is malicious behavior of malware in android platform.

PO3 to evaluate the performance of classifier technique, by look at the parameter that has been choose such as speed, accuracy, and the result by percentage.

1.6 Scope

For the project scope, it will divide by 3 parts which is type of malware, software and hardware that used in this project.

1. Case study
    i. Type of malware

    To perform the data collection on this project, it will require a malware with different variant. Total malware application has been collect are 75 and clean application also 75. For the clean application will be randomly chooses at market place or any resources. The reason following malware has been choose that is because this type of malware requires only 4 minutes to run all malicious code inside Smartphone for example within 4 minute this type of

malware will try to connect to the remote host to send privacy data. The family of malware as followed.

i. Droidkungfu1 family 15 application
ii. Droidkungfu4 family 15 application
iii. BaseBridge family 10 application
iv. Goldream family 10 application
v. Genimi family 8 application
vi. Droidkungfu2 family 12 application
vii. Bindbot family 5 application

ii. Duration of data collection

On this project, for data collection it will take approximately ten minute for each malware on client side. For ten minute it has a several procedure and process. The reason we choose ten minute for each application it is because, ten minute is enough to capture the behavior of malware. For this project 2 month are required to collect the data.

iii. Algorithm Model

The algorithm that has been chooses for classifier techniques to classify the behavior of malware are as followed:

i. Trees.J48
ii. Trees.RandomForest
iii. Rules.PART
iv. Function.Logistic
v. Meta.MultilClassClassifier

This is algorithm that has been chosen and compares to produce the output and result. This is because these 5 algorithms show the best performance and high accuracy for android data. Then, the percentage or correctly classified instance is 90% and above.

2. Software

For the software on this project, we use the android Operating system version 4.2, because android is most popular among us and the application easy to download.

3. Hardware

On this project for the client it will use Samsung galaxy tab GT p6800. This device use during data collection

1.7 Expected Output

This proposes project has its significance and also the advantages of this paper.

1. Towards body of knowledge
   a. Theoretical
      i. Performance indicator.

      The end of this project we hope can produce the performance indicator to use to evaluate the performance of classification technique for android malware data.

      ii. Malware Behavior

      After pre-processing we expected can provide list of the malware behavior that created by application with malware.

      iii. The Best Classification Technique

      After of all phase we expected that this project can provide the best and most accuracy algorithm to used for other researcher to classify android data.

   b. Practical
      i. Data Collection

      The end of the project we expect that it can provide a formal procedure to perform data collection for others researcher.