

**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)

No 26 jalan Selaseh 3,
Alamat tetap: _____

Taman Selaseh Fasa 1,

68100 Batu Caves, Selangor

Tarikh _____

(TANDATANGAN PENYELIA)

PN. SYARULNAZIAH BINTI
ANAWAR

Tarikh: _____

CATATAN: * Tesis dimaksudkan sebagai Laporan Projek Sarjana Muda (PSM).

** Jika tesis ini SULIT 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 pengelasan. 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 segala masalah dan persoalan yang ada di mana 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

CHAPTER	SUBJECT	PAGE
	DEDICATION	i
	ACKNOWLEDGMENT	ii
	ABSTRACT	iii
	ABSTRAK	iv
	TABLE OF CONTENTS	v
	LIST OF TABLES	xi
	LIST OF FIGURES	xiii
1	INTRODUCTION	
	1.1 Introduction	1
	1.2 Project Background	1
	1.3 Problem Statement	2
	1.4 Project Question	2
	1.5 Project Objective	3
	1.6 Scope	3
	1.7 Expected Output	5
	1.8 Report Organization	6
	1.9 Summary	7
2	LITERATURE REVIEW	
	2.1 Literature Review	8
	2.2 Fact and Finding	8

2.2.1	Classifier in Machine learning	9
2.2.1.1	Machine Learning	9
2.2.1.2	Classifier	10
2.2.1.3	Type of Machine Learning Technique and Algorithm	11
2.2.1.3.1	Bayes	11
2.2.1.3.2	Decision Tree	12
2.2.1.3.3	Function	13
2.2.1.3.4	Lazy	14
2.2.1.3.5	Meta	15
2.2.1.3.6	MISC	15
2.2.1.3.7	Rules	16
2.2.2	Previous Work	16
2.2.3	Previous Work	18
2.2.3.1	Experimental Results	19
2.2.4	Conclusion of Previous Work	20
2.2.5	Malware Detection Technique	20
2.2.5.1	Anomaly Detection (Based Detection)	22
2.2.5.2	Specification Based Detection	24
2.2.5.3	Signature-based detection	25
2.2.5.4	Malware	26
2.2.6	Type of Malware and Behavior	28
2.2.6.1	Viruses	28
2.2.6.2	Worms	28
2.2.6.3	Trojan Horse	29
2.2.6.4	Backdoors	29
2.2.6.5	Spyware	29
2.2.6.6	Rootkits	29

2.3	Performance of Indicators	30
2.3.1	Accuracy	30
2.3.2	Correctly and Incorrectly Classified Instance	30
2.3.3	Mean Absolute, Error Kappa Statistic	31
2.3.4	Average True Positive and Average False Positive	31
2.3.5	Speed	31
2.4	Summary	32
3	RESEARCH METHODOLOGY AND DESIGN	
3.1	Research Methodology and Design	33
3.1.1	Type of research	33
3.1.2	Quantitative Methodology	33
3.1.3	Rationale of Selected Methodology	34
3.2	Research Framework	34
3.2.1	Methodology of Project	34
	3.2.1.1 Literature Review	35
	3.2.1.2 Data Collection	35
	3.2.1.3 Data Validation	35
	3.2.1.4 Pre-processing	35
	3.2.1.5 Classification	35
	3.2.1.6 Evaluate	36
3.2.2	Experimental Setup	36
3.3	Hardware Requirements	37
3.3.1	Smartphone with Android Platform	38
3.3.2	Server	38
3.3.3	Service Provider	39
3.4	Software Requirement	40
3.5	Research Process	41
3.5.1	Server Component	42
	3.5.1.1 Storage	42

	3.5.1.2 Malware	42
	3.5.2 Package Extractor	44
3.6	Project Schedule and Milestones	45
	3.4.1 Milestone	46
3.7	Summary	47
4	DATA COLLECTION AND VALIDATION	
4.1	Data Collection and Validation	48
4.2	Data collection	49
	4.2.1 Initial Configuration	50
	4.2.1.1 Run Script	52
	4.2.1.2 Application Installation	54
	4.2.1.3 Explore the Application	54
	4.2.1.4 Browsing Internet	55
	4.2.1.5 Messaging With Other Client	55
4.3	Validation	56
4.4	Data Collection and Validation process “Step By Step”	56
	4.4.1 Server Part	56
	4.4.2 Client Part	59
4.5	Data Validation	65
4.6	Summary	66
5	PRE-PROCESSING	
5.1	Pre-Processing	67
5.2	Software Requirement for Pre-Processing	67
	5.2.1 Python	67
	5.2.2 Weka	68
5.3	Major task in Data Pre-processing	68
	5.3.1 Data Cleaning	68

5.3.2	Data Transformation	68
5.3.3	Data reduction	69
5.4	Step of Pre-Processing	70
5.5	Command to Calculate Syscall on Excel	76
5.6	Syscall Process	77
5.7	Pre-processing On Weka	78
5.7.1	Load File inside Weka	78
5.7.2	Select attribute	79
5.7.2.1	Attribute Evaluator to CfsSubsetEval	79
5.7.2.2	Searching Method	80
5.7.3	Testing to choose Numbers Attribute	
	Fold Percentage	81
5.7.4	Comparison between Two Attribute Selection	
	Mode	82
5.7.4.1	Attribute selection mode Use full	
	Training mode	83
5.7.4.2	Attribute selection mode “Cross Validation”	83
5.8	Malware behaviors	85
5.9	Summary	87
6	Analysis and Result	
6.1	Analysis and Result	88
6.2	Evaluation of Classifier Performance	88
6.2.1	Cross-validation	89
6.3	Algorithm Has Been Tested	89
6.4	Analysis	92
6.4.1	Trees.J48	92
6.4.1.1	Discussion Result J48	93
6.4.2	Trees.RandomForest	94
6.4.2.1	Discussion RandomForest Result	94

6.4.3	Rules. PART	95
6.4.3.1	Discussion PART Result	96
6.4.4	MultiClassClassifier	97
6.4.4.1	Discussion MultiClassClassifier Result	97
6.4.5	Function.Logistic	98
6.4.5.1	Discussion Logistic Result	99
6.5	Comparison between Five Algorithm	100
6.6	Result from Analysis	102
6.7	Validation using T Test	103
6.7.1	Cross-Validation for Accuracy	104
6.7.2	Result Accuracy for T Test	106
6.7.3	Cross-Validation for Speed	106
6.7.4	Result Speed T Test	109
6.8	Result T Test	109
6.9	Validation Different Data Set	109
6.10	Summary	110
7	CONCLUSION	
7.1	Introduction	111
7.2	Research Summarization	111
7.3	Contribution of the Research	112
7.4	Project Limitation of the Research	112
7.5	Project Future	113
7.6	Summary	113
	REFERENCES	114
	APPEDICES	116

LIST OF TABLES

TABLE	TITLE	PAGE
1.1	Problem Statement	2
1.2	Project Question	2
1.3	Project Objective	3
2.1	Result for Different Algorithm in Classifier Technique (Othman& shan Yau, 2007)	17
2.2	Data for the project	19
2.3	Result from the project	19
2.4	Static and Dynamic Advantages and disadvantages	23
2.5	Type of Malware Attack (Nour Saffaf, 2009)	27
3.1	Smartphone Specification	38
3.2	Server Specification	39
3.3	Access Point Specification	39
3.4	Gantt Chart	45
3.5	PSM 1 Milestones	46
6.1	38 Algorithm Has Been Tested	90
6.2	J48 Result	92
6.3	J48 Calculation Of accuracy	93
6.4	RandomForest Result	94
6.5	RandomForest Calculation of Accuracy	95
6.6	Part Result	95
6.7	Part Calculation of Accuracy	96
6.8	MultiClassClassifier Result	97
6.9	MultiClassClassifier Calculation of Accuracy	98

6.10	Logistic Result	98
6.11	Logistic Calculation of Accuracy	99
6.12	Result from Five Algorithm	100
6.13	The Sample of Data Correctly Classified Instance %	104
6.14	Description Statistic	105
6.15	t-Test Two-Sample Assuming Unequal Variances	105
6.16	The Sample of Data Time Taken (second)	107
6.17	Description Statistic	107
6.18	t-Test: Two-Sample Assuming Unequal Variances	108
6.19	Result of Classifier for Validation	109

LIST OF FIGURES

FIGURE	TITLE	PAGE
2.1	Process Classifier (Xin Guan, 2013)	10
2.2	Decision Tree Example	12
2.3	Bar Chart of Comparison between Parameters (Othman& shan Yau, 2007)	17
2.4	Methodology of the Face Recognition Project	18
2.5	Detection Tree (Idika& Mathur, 2007)	25
3.1	Methodology for Android Malware Detection through Machine Learning Analysis	34
3.2	Physical Design	36
3.3	Logical Design	37
3.4	Architecture for Android Malware Detection Trough Machine Learning Analysis	41
3.5	Server Component Diagram	42
3.6	Client Component Diagram	44
3.7	Gantt Chart for This Project	46
4.1	Data Collection Procedure	49
4.2	How User Linux Kernel Executed (Burguera& Zurutuza, 2011)	50
4.3	Initial Configuration Flow Chart	51
4.4	Cronroot Command	52
4.5	Cronrfold Command	52
4.6	Command File Script	53
4.7	Runst Command	53
4.8	Runstop command	54

4.9	Data Collection Procedure In minutes	55
4.10	Apktool On CMD	57
4.11	File After Extract by apktool	57
4.12	All folder and file has been create by apktool	58
4.13	Package Name of Malware	58
4.14	Create Syslink To system File	59
4.15	Edit Runst File	60
4.16	Edit File Runst	60
4.17	Run Script	61
4.18	Install Application	61
4.19	Open Application	62
4.20	Game main page	62
4.21	Force Stop Application	63
4.22	Run Runstop Script	63
4.23	Copy to ExtSdCard	64
4.24	Create Folder	64
4.25	Cut File Has Been Create By Script	65
4.26	Example Application Stop	66
5.1	Test No of Fold for Attribute Selection	81
5.2	All File Inside Python Folder	70
5.3	Example Strace File	70
5.4	This is Data Inside Strace File	71
5.5	Copy Strace File to Python Directory	71
5.6	Command Prompt with Python Directory as Path	71
5.7	Command To Convert To .csv Fill	72
5.8	Finish Convert To csv File	72
5.9	Example Error Data	73
5.10	Output File before Rename	73
5.11	Output File after Rename to .csv	73
5.12	Row Data on Excel	74
5.13	Sort the Row Data in Excel	74
5.14	After Sort the Data	75

5.15	To Identify All Command on .csv file	75
5.16	Process after Process Has Been Calculated	76
5.17	Data Set for Classifier	77
5.18	Load csv File into Weka	78
5.19	After load csv file	79
5.20	Select attribute process	81
5.21	Graph Result Comparison between 4 No Attribute Fold Percentages	82
5.22	Result from “Use full training mode”	83
5.23	Result from cross validation mode	83
5.24	Percentage of effeteness for each attributes	84
5.25	Before Remove attribute that not affect substantially	84
5.26	After Remove attribute that not affect substantially	85
5.27	Malware Behaviour	86
5.28	Classification Visual Tree	86
6.1	Comparison between Accuracy and Tike Taken for all Algorithm	91
6.2	Comparison Correctly and Incorrectly Classified Instances between 5 Algorithms	100
6.3	Comparison of Kappa Statistic between 5 Algorithms	101
6.4	Comparison of Average TP and Average FP between 5 Algorithms	101
6.5	Comparison of Time Taken between 5 Algorithms	102
6.6	T Test Graph for Accuracy	106
6.7	T Test Graph for Accuracy	108

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 classification 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.