DIMENSION REDUCTION FOR CLASSIFICATION USING PRINCIPAL COMPONENT ANALYSIS (PCA) TO DETECT MALICIOUS EXECUTABLES

ZALIFAH BINTI JAMAL

MASTER OF COMPUTER SCIENCE
(INTERNETWORKING TECHNOLOGY)

2017

Faculty of Information and Communication Technology

DIMENSION REDUCTION FOR CLASSIFICATION USING PRINCIPAL COMPONENT ANALYSIS (PCA) TO DETECT MALICIOUS EXECUTABLES

Zalifah Binti Jamal

Master of Computer Science (Internetworking Technology)

2017

DIMENSION REDUCTION FOR CLASSIFICATION USING PRINCIPAL COMPONENT ANALYSIS (PCA) TO DETECT MALICIOUS EXECUTABLES

ZALIFAH BINTI JAMAL

A dissertation submitted
in fulfillment of the requirements for the degree of Master in Computer Science (Internetworking Technology)

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2017

DECLARATION

I declare that this dissertation entitled “Dimension Reduction for Classification Using Principal Component Analysis (PCA) To Detect Malicious Executables” is the result of my own research except as cited in the references. The dissertation has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:	……..
Name	:	Zalifah Binti Jamal
Date	:	……..……

APPROVAL

I hereby declare that I have read this dissertation and in my opinion this dissertation is sufficient in terms of scope and quality as a partial fulfillment of Master of Computer Science (Internetworking Technology).

Signature	:	……..
Supervisor Name	:	Professor Dr. Burairah Bin Hussin
Date	:	……..……

DEDICATION

This thesis is dedicated to my lovely parents, my husband and my children who support me all the way since the beginning of my studies.

ABSTRACT

The development of cyberspace is not only facilitate people's lives. It should also be in line with security awareness related to personal and enterprise systems. Estimates of the number of new malware in 2013 reached 600 million, and has grown rapidly in recent years. Malware can attack a wide variety of computing devices and mobile devices are no exception. The number of malware attacks this execution on a large scale. This is a big challenge for malware detector. There are several ways of classification that are used to verify the accuracy of the research. Most classifiers have too many combinations that are difficult to assess, change often (optimal) and should get a brief training period. This study is aimed at reducing high-dimensional vector space to a lower dimension, thus reducing the problem of lack of accuracy of results. This study used a new approach, namely the Principal Component Analysis (PCA). PCA will make a classification so that the process can be done automatically and efficiently. PCA can reduce the number of dimensions of space by extracting features that describe the data set so that data sets can be confirmed precisely as if the entire data set together. The purpose of this study to investigate the malware will be selected, reducing the dimensions of the model that will be used to detect malware and to validate the models to find a minimum set of data to detect malware data. In order to find the right combination of features and options classification, two different sets of selection criteria used by two machine learning classifier. Result classification was assessed using the True Positive Rate (TPR), the false negative rate (FNR) and the accuracy of the feature selection approaching or exceeding 95% accuracy.

ABSTRAK

Pembangunan ruang siber bukan sahaja memudahkan kehidupan orang ramai. Ia juga perlu seiring dengan kesedaran keselamatan yang berkaitan dengan sistem peribadi dan korporat. Anggaran bilangan malware baru pada tahun 2013 mencapai 600 juta, dan ia telah berkembang dengan pesat dalam tahun-tahun kebelakangan ini. malware ini boleh menyerang pelbagai peranti komputer dan peranti mudah alih adalah tidak terkecuali. Jumlah serangan malware pelaksanaan ini adalah dalam skala yang besar. Ia adalah satu cabaran besar untuk malware pengesan. Terdapat beberapa cara pengelasan yang digunakan untuk mengesahkan ketepatan kajian. Kebanyakan pengelasan mempunyai terlalu banyak kombinasi yang sukar untuk menilai, perubahan yang kerap (optimum) dan terpaksa mendapatkan tempoh latihan yang singkat. Kajian ini bertujuan untuk mengurangkan ruang vektor dimensi yang tinggi kepada dimensi yang lebih rendah, sekali gus mengurangkan masalah kekurangan ketepatan keputusan. Kajian ini menggunakan pendekatan baru iaitu Analisis Komponen Utama (PCA). PCA akan membuat klasifikasi supaya proses itu boleh dilakukan secara automatik dan cekap. PCA boleh mengurangkan bilangan dimensi ruang dengan mengekstrak ciri-ciri yang menerangkan set data supaya set data yang boleh disahkan dengan tepat seolah-olah keseluruhan dataset disatukan bersama.. Objektif kajian ini untuk menyiasat malware akan dipilih, mengurangkan dimensi untuk model yang akan digunakan untuk mengesan malware dan untuk mengesahkan model untuk mencari satu set minimum data untuk mengesan data malware. Dalam usaha untuk mencari gabungan ciri dan pilihan klasifikasi, 2 set kriteria pemilihan yang digunakan oleh 2 penjodoh pembelajaran mesin. Keputusan klasifikasi dinilai dengan menggunakan Kadar Positif Yang Benar (TPR), Kadar palsu negatif (FNR) dan ketepatan pemilihan ciri yang menghampiri atau melebihi 95% ketepatan.

ACKNOWLEDGEMENTS

In the name of Allah, Most gracious, Most Merciful. All praises belongs to Allah. First of all, I would like to thank to Allah Al the Mighty, who made me capable to complete the thesis throughout those difficult years.

First and foremost, I would like to thank to my supervisor, Professor Dr. Burairah Bin Hussin for his excellent supervision, guidance, supporting and encouragement towards in completing my thesis. May Allah reward him with a reply that much better than what all he has done.

I am in debt and owe great thanks to my beloved husband Mazli Basiran, my mother, my siblings and my dearest children Nurin Ezatty, Qaisara Faqeehah, Muhammed Baseer Rifa'e and Abdurrahaman Auff for their patience, inspiration, continuous encouragement and thoughtful advice throughout my years as a Master student.

The dataset used in this study was given by Mr. Mohd Zaki Mas'ud, lecturer from the University of Technical Malaysia Melaka. I would like to express appreciation to him. Last but not least, my special thanks to all my friends for their time, understanding, advice and continues moral support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.0 Background and Context

When computer users surf the internet for indirect data downloaded it along with malicious software. It's called download attack. The new malware is growing by a staggering rate. . The existence of thousands of malicious code in cyberspace has been rampant. Among the cyber-attack as it is contained in the malicious emails, files and even malicious attachment contained in Thumb drive within one day Microsoft received more than 150 new files were not known if the file has been infected with malware or not. The file is in the form of a very large scale. These files need to be analyzed to determine the types of malware and so do the classification according to type of malware. Thus an analysis manually is impossible because of the size of those files are in a very large scale (Dahl et al. 2013). In the period of January 2013 to December 2014 there were approximately 60 million malware has been reported. Malware programmers have also evolved and they have the expertise in creating new variants of malware by creating a variety of techniques such as the interception techniques, encryption, polymorphism and it's very confusing. This is the most serious threat to the global security.

Refer to (Rieck et al. 2011) study Malware, known as malware, poses a significant threat to the security of computer systems. The variants of the number and variety of classic invalid security defences allow the Internet host to present millions of computer viruses,

internet worms and trojans infected with malware. While malware confusion and polymorphism detection are used at the file level for large-scale barriers, runtime analysis of malware binaries and characterization of malware risks provides a means of protection.

Large amounts of data will delay the process of training data, making it difficult for the test data. This situation had a negative impact on the performance of a computer system to implement malware detection. Therefore, to overcome this problem it goes a good system to ensure that the rate of detection is high. (Nziga and Cannady 2012). Reducing the amount of data to a minimum set is very important to analyze the data easily and quickly.

According to (Nziga and Cannady 2012), internet traffic is monitored by Network Intrusion Detection Systems (NID) to detect malware. NID task is to check the traffic in and out of traffic to identify suspicious activity. The challenges they encounter is when faced with how to analyze the activities that deviate from the norm.

According to the study by (Dahl et al. 2013) Malware classification system is based on a set of binary features. In the analysis of malware analysts use the binary string rarely based files, application programming interface (API) tri-gram and API calls plus the value of the parameter. Feature selection must be done in advance to ensure the best accuracy. Logistic regression of all features will produce reasonable accuracy on a large scale dataset. However, they produce a lot of errors. Classification is automatically required to solve this problem.

Anti-Virus vendor's task is to figure out how to prevent the development of malware, so do not spread because they are faced with tens of thousands of malware every day. When you see this situation as Anti-Virus vendors endless war with cybercriminals. (Jang et al.

2015a). Data storages cannot accommodate the number of properties malware that has been extracted into it and cause processing is too complex. (Mas'Ud et al. 2014a).

Dimension reduction means finding a subset of the original features as possible, i.e. removing irrelevant data, excessive or noisy data. This feature should be removed because it is not important data and this data is misleading analysts Data will be too complex Dimension reduction can improve the classification of data and it has proven effective. Apart from the classification algorithm is rapid and improve the comprehensibility of the results. When faced with high dimensional data dimension reduction is an important issue. Therefore do dimension reduction before performing classification tasks. (Zhang and Jing, n.d.)

Antivirus vendors are working hard to detect and classify malware detection performance effectively and efficiently. This problem can be solved by using signal processing and machine learning methods to detect new malware automatically.

1.1 Problem Statement

The rise of the virtual world not only facilitate people's lives but also increase safety awareness related to personal and corporate systems. In addition, the estimated number of new malware in 2013 reached 600 million and it has grown exponentially in recent years. The amount of malicious data execution must be too large. This condition causes the detector to face many challenges to make the detection of malware in cyber technology. Most malware researchers focused on data mining to detect unknown malware.

There are various methods of classification are used to verify the accuracy of the study. Most of these methods have a problem as a matter of selecting the features. Too many combinations that are difficult to assess, frequent changes (optimum) and had to get a shorter

training time. In addition, reducing high-dimensional vector space to a 5th dimensional information in data compromise and reduce the loss of accuracy results.

The identification process would take a long time if it is done at an early stage. Identification of the data will be more easily and automatically if the classification is done at an early stage. We do not need to compare data one by one because it takes a long time to complete.

Classify patterns, which aims to reduce the number of space dimensions subspace by extracting features that describe the data set so that data can be confirmed accurately and quickly without the need to run one after the recognition process.

Thus, in this study, the main subject of concern will focus on the investigating feature selection for detection of malicious code, applying dimension reduction approach for obtaining a model use to malicious detection and verifying the model for malicious detection.

1.2 Motivation of Study

The identification process would take a long time if it is done at an early stage. Identification of the data will be more easily and automatically if the classification is done at an early stage. We do not need to compare data one by one because it takes a long time to complete. There is a new approach using PCA, it will create a classification so that the process can be done automatically and quickly. Classify patterns, which aims to reduce the number of space dimensions subspace by extracting features that describe the data set so that data can be confirmed accurately and quickly without the need to run one after the recognition process. In addition, principal component analysis is analysing data to identify

patterns. The results obtained from it can reduce dimensional dataset with minimal loss of information Principal Component Analysis will cause the highlighting feature space to a smaller subspace.

1.3 Objective of Study

The following objectives should be satisfied in order to achieve the goal:

1. To investigate feature selection for detection of malicious code.
2. To apply dimension reduction approach for obtaining a model use to malicious detection.
3. To verify the model for malicious detection.

1.4 Question of Study

The aim of this study is to better understanding and concern with:

i. How to investigate feature selection for detection of malicious code.
ii. How to apply dimension reduction approach for obtaining a model use to malicious detection.
iii. How to verify the model for malicious detection.

1.5 Scope of Study

The scope of this study is to detect and analyze the existence of malicious data, this study selected two sets of characteristics. Feature sets of system calls, and feature sets based on feature sets of system calls selected by the choice method of principal component analysis for dimension reduction techniques.

CHAPTER 2

LITERATURE REVIEW

2.0 Introduction

This paper present a review of literature pertinent to the study as presented by various researchers, scholar, analyst and authors. This chapter summaries literature that has been reviewed and will be reviewed for the purpose of the study which is Dimension Reduction for Classification Using Principal Component Analysis (PCA) To Detect Malicious Executables. The literature covers theoretical framework and an overview of the literature of previous studies, findings and recommendation showing the study gap to be filled. The conceptual framework of the study and summary are provided.

2.1 The growth of Malicious Executables

Currently, high-growth processing technology has led to an exponential growth in the data harvested with respect to two dimensions and size of the sample. These growth trends UCI machine learning repository. Effective and efficient management of data is becoming increasingly demanding. This traditional manual management dataset does not seem practical. Thus techniques such as machine learning and data mining have been developed to automatically get recognition and recognize patterns of data.(Shafreen Banu and Hari Ganesh 2015).

Refer to (Pirscoveanu et al. 2015) study, the use of the Internet trend exponentially over the past few years, with the modern society more and more depending on global communications. Simultaneously, the Internet is increasingly being used by criminals, a large black there have been hackers or other criminals in the market Intent can purchase malware or use malicious services rental fees. This provides a powerful incentive for hackers modify and increase the complexity of malicious code in order to improve confusion to reduce opportunities detection by antiviral program. This results in multiplexing fork or the same type of malicious new implementation software may spread out of control. Based on AV testing, About 390,000 new malware samples were registered every day, this creates a problem of handling large amounts of unstructured data from malware analysis. This makes the antivirus vendors face challenges detect zero-day attacks and publish updates in a reasonable way time frame to prevent infection and reproduction.

According to (Akhuseyinoglu and Akhuseyinoglu, n.d.), in their study, when a user downloads an Android application to the phone, the list of permissions is introduced by the application during installation and running. Android's licensing system gives users security risks and provides a free choice to determine the security permissions of the application. However, users typically do not have sufficient capacity to define the access rights, thus giving the necessary guidance. A user-centric approach is challenging and can motivate users to receive any licenses required by applications that cause security risks hackers have a variety of subtle ways to hurt mobile devices and trick users into discovering new ones every day. As a counter attack to this growing army, the security solutions proposed by researchers have been increasing.

In recent years, unauthorized access to remote information has increased significantly and has become a threat to many organizations dealing with security and insecurity data uses emerging technologies, such as cloud computing.(Suthaharan and Panchagnula 2012).

According to (Nziga and Cannady 2012) Network Intrusion Detection Systems (NIDS), but including service attacks, port scans, and try to get the additional rights and denial of access to unauthorized users on the network is not limited to Internet traffic monitoring to detect malicious activity. NIDS find suspicious patterns of all incoming packets, outgoing or check local traffic. The challenge is greater data network Intrusion Detection Systems (NIDS) is analyzed to identify the actions that deviate from normal behaviour.

Subsequently, the extracted features provide plenty of storage for collecting and processing can contribute to an increase in complexity, so it's only storage, memory, CPU and power consumption on a mobile device is not a favorable option. Choose a nice feature when the pre-processing of data more efficient to identify the infected Android application will be processed on the machine learning classifier. Yet while preserving the accuracy of cuts to selection (Mas'Ud et al. 2014b).

(Pirscoveanu et al. 2015) said the challenge of feature reduction is minimization the number of features without losing performance classification. Function of personalized reduction purposes limit the final number of features within a combination matrix.

Refer to this study, data mining methods in a large number of data patterns, and these patterns can be used to find similar data for future events. One of the main problems of data mining methods to detect new malicious programs are the choice of using the feature. This process is tackled from two different angles from viruses optimal signatures in a data

extraction, and using a complex classification schemes to identify the more general feature (Lai 2008).

According to (Lai 2008) too, several approaches are capable of detecting unknown malicious activity from detection. Some degree of success in machine learning or data mining methods can detect new or unknown malicious functions. Amenities successfully detect malicious functions to apply data mining or machine learning is a key. The researchers propose a method to extract features of the most representative characteristics of the virus. They are our assortment of strings, based on this technology reaches high detection rates and show that you cannot expect to do the same in real world conditions.

2.2 Machine Learning

(Author 2016) Machine Learning structured and unstructured data plays an increasingly important role in the process. Document clustering Committee unstructured data (text documents) is used as the basis of its content, it is a popular machine learning technique for data analysis is to understand the patterns. Text mining and unstructured data collection (K- means) will be transformed into structured data using techniques stages semi-structured data. Corporate Banking, financial services and insurance industry "to sell and profit opportunity for fraud detection and identity" that can be used for events, such as machine learning technique to another. The researchers combined the use cases in the industry, both in the implementation of the algorithm for document clustering and classification algorithms (Decision Tree, Random Forest and naive Bayes) focuses on a set. Moreover, as a result the performance of the three classification methods, "accuracy", "accurate", and "Return" allows us to calculate performance measures "Confusion Matrix" will be compared by calculation.