ANOMALY DETECTION USING K-MEANS CLUSTERING AND DECISION TREE CLASSIFICATION

HAZLIN BINTI MOHD ZIN

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS TESIS*

JUDUL: ANOMALY DETECTION USING K-MEANS CLUSTERING AND DECISION TREE CLASSIFICATION

SESI PENGAJIAN: 2015/2016

Saya HAZLIN BINTI MOHD ZIN

mengaku membenarkan tesis (PSM/Sarjana/Doktor Falsafah) 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: 179-Y-Jalan Sek Keb Lundang, 15150, Kota Bharu, Kelantan	(Dr S.M. Warusia Mohamed Bin S.M.M Yassin)
Tarikh : 25/08/2016	Tarikh : 25/08/2016

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

ANOMALY DETECTION USING K-MEANS CLUSTERING AND DECISION TREE CLASSIFICATION

HAZLIN BINTI MOHD ZIN

This report is submitted in partial fulfilment of the requirement for the Bachelor of Computer Science (Computer Security) With Honours

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2016

DECLARATION

I hereby declare that this project report entitled

ANOMALY DETECTION USING K-MEANS CLUSTERING AND DECISION TREE CLASSIFICATION

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

STUDENT : ______________ Date: 25/8/16

(HAZLIN BINTI MOHD ZIN)

I hereby declare that I have read this project and found

this project is sufficient in term of the scope and quality for the award of

Bachelor of Computer Science (Computer Security) With Honours.

SUPERVISOR : ______________ Date: 25/8/16

(DR. S.M. WARUSIA MOHAMED BIN S.M.M YASSIN)

DEDICATION

To ALLAH SWT, my beloved parents, family, supervisor, friends and myself.

ACKNOWLEDGEMENT

I would like to express my thank and appreciation to my supervisor Dr SM Warusia Mohamed SMM Yassin for his continues encouragement, advices and guidance throughout this research as the freedom that he give me while I was working on my research and their brilliant to new idea.

Special thanks to my friends for their support, motivation and always give help to share ideas and knowledge as I need them. Hopefully we are going to grab succeed together.

Last but not least, my sweetest appreciation to my parents and family for their continuous support, patience and encouragement to me. Their prayers and good wished will always help me to keep strong during my difficult times. I am grateful and thankful to them.

ABSTRACT

Nowadays, our country Malaysia was not exempt from cyber incidents. Although there are many types of security methods like access control, encryption, firewall are used but network security breaches increase day by day. With such unpredictable pattern of attacks, our defense calls for an urgent need to efficiently identify attacks and to classify them based on the degree of threats that they pose. One of the components of security that suit the 'defense in depth' model is called the Intrusion Detection System (IDS). IDS become an important defense to block for any network intrusion. An IDS is capable of detecting and sending early alarm upon risk exposure caused by any attack. A growing interest in the investigation of anomaly detection sparks from the ability of the approach to detect unknown attacks and to evaluate. A new hybrid mining approach is to improving current anomaly detection capabilities in IDS that would be securing an information infrastructure which is K-means clustering method and classification method. Thus, an urgent action needed to detect any attacks effectively. Data mining is the latest technology that been introduced in network security to fine regularities and irregularities in large data set. In this project, a hybrid data mining approach formed by combining the K-means clustering and classification. For the accuracy result, the detection and false alarm rate will be compared to the previous techniques that have been done before on the related research.

ABSTRAK

Pada masa kini, negara kita Malaysia tidak terkecuali daripada insiden siber. Walaupun terdapat banyak jenis kaedah keselamatan seperti kawalan akses, *encryption*, *firewall* digunakan tetapi pelanggaran keselamatan rangkaian meningkat hari demi hari. Dengan corak serangan yang tidak menentu, pertahan siber Negara ini terpanggil untuk untuk mengenal pasti jenis-jenis serangan dan untuk mengklasifikasikan mereka berdasarkan kepada tahap ancaman yang mereka menimbulkan. Salah satu komponen keselamatan yang sesuai dengan *defense in depth* model dipanggil Sistem Pengesanan Pencerobohan (IDS). IDS menjadi pertahanan penting untuk menyekat apa-apa pencerobohan rangkaian. IDS mampu mengesan dan menghantar penggera awal apabila terdedah risiko yang disebabkan oleh mana-mana serangan. Daripada kejadian-kejadian ini, timbul minat untuk melakukan penyisatan terhadap *anomaly detection* yang mana ianya berkebolehan untuk mengesan pencerobohan dan melakukan penyelidikan. Pendekatan yang baru iaitu *hybrid mining* adalah untuk meningkatkan keupayaan pengesanan anomali semasa di IDS yang akan mendapatkan infrastruktur maklumat yang *K-means* dan *classification*. Oleh itu, tindakan segera diperlukan untuk mengesan sebarang serangan berkesan. *Data mining* adalah teknologi terkini yang diperkenalkan dalam keselamatan rangkaian denda kebiasaan dan penyelewengan dalam set data yang besar. Dalam projek ini, pendekatan *hybrid data mining* dibentuk dengan menggabungkan *K-means* dan *classification*. Untuk keputusan ketepatan, pengesanan dan kadar penggera palsu akan dibandingkan dengan teknik sebelumnya yang telah dilakukan sebelum ini kepada penyelidikan yang berkaitan.

TABLE OF CONTENTS

CHAPTER II LITERATURE REVIEW

CHAPTER III PROJECT METHODOLOGY

CHAPTER IV DESIGN

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1. INTRODUCTION

Computer security has turned into a need because of the bulk of information technologies in our daily life. The mass usage of computerized system has given an ascent to risky threats such as zero-day vulnerabilities, mobile threats and etc. Nowadays, our country Malaysia was not exempt from cyber incidents. Although there are many types of security methods like access control, encryption, firewall are used but network security breaches increase day by day. From the summary report made by Malaysian Computer Emergency Response Team, each of every year, they handled a large number of incidents, for example, intrusion, fraud, cyber bullies and others. Any computer that is either connected through physical or in a wireless environment is bared to the risk threat.

With such unpredictable pattern of attacks, our defense calls for an urgent need to efficiently identify attacks and to classify them based on the degree of threats that they pose. One of the segmentation in the security that suit the 'defense in depth' model is called the Intrusion Detection System (IDS). IDS turn into an imperative defense to hinder for any network intrusion. An IDS is capable of detecting and sending early alarm upon danger exposure caused by any attack. This wills give an alert to the system administrators to take some action corresponding response measurements; along these it will reduce the possibility of bigger losses.

An enthusiasm in the investigation of anomaly detection sparks from the ability of the approach to detect unknown attacks and to evaluate them in term of accuracy, detection rate and false alarm rate. A new hybrid mining approach is to enhancing current anomaly detection capabilities in IDS that would be securing an information infrastructure which is K-means clustering method and classification method. Data mining approach is used as a detection method to discover unforeseen attack. Thus, an urgent action needed to detect any attacks effectively. Data mining is the most recent technology that been acquainted in network security to fine regularities and irregularities in large data set. In this project, a hybrid data mining approach formed by combining the K-means clustering and Decision Tree classification. For the accuracy result, the detection and false alarm rate will be compared to the previous techniques that have been done before on the related research.

1.2. PROJECT MOTIVATION

A cyber attack is a planned misuse of PC frameworks, development of innovation subordinate, and network. Cyber attack use malicious code to alter PC code, justification or data, bringing about problematic results that can exchange off information and lead to cyber crimes, for example, information and fraud. Cyber attack is socially or politically induced assaults brought out basically through the Internet. Attack will concentrate on the overall population or national and corporate associations and are brought out through the spread of malicious programs (viruses), unapproved web access, fake sites, and different method for taking individual or institutional data from focuses of attacks, bringing about extensive harm.

Since cyber attack has turned into a novel weapon of war and their industriousness makes it basic to coordinate more exact Intrusion Detection System (IDS) that skilled to boosting accurately the information, erroneously information and decrease the recognition time to distinguish the attack. Anomaly based detection system which employ K-means clustering and classification method which is a significant field to explore the above capabilities. For the continuous enhancement of intrusion detection capabilities, the detection of unforeseen attack and the approach is the motivation for me to do this research.

1.3. PROBLEM STATEMENT (PS)

An intrusion detection system (IDS) monitors network traffic and monitors for suspicious action and alarms the system or network administrator. In some cases the IDS may also react to anomalous or malicious traffic by taking action for example obstructing the client or source IP address from getting access to the network. IDS come in a variety of flavors and approach the objective of detecting suspicious traffic in various ways.

However, anomaly detection has their own particular strength over signature-based engines in that a new attack for which a signature does not exist can be detected if it falls out of the normal traffic patterns. Anomaly detection is frequently associated with high false alarm with just moderate accuracy of detection rates.

Therefore, there is a requirement for an approach that could detect and identify such unforeseen attacks. A hybrid mining approach is proposed through of K-Means clustering with a variant of *k*th cluster center and with various number classifiers so as to diminish the false alarm. K-Means clustering is an anomaly detection technique that partitions data into corresponding group called clusters, whereby all data in the same cluster are similar to each other. Hybrid approach will be clustering all data into the corresponding group before applying a classifier for classification purposes.

A classifier is a machine learning approach where the learned attribute is categorical. It is used after the learning process to classify new data by giving them the best target of prediction.

This approach could perform better in term of to achieve the best possible rate of false alarm for unforeseen attack remain as challenging task and unresolved issues such as predicting an intrusion as normal instances and normal instances as attacks or intrusion become inevitable limit in building effective anomaly detection. Thus, the problem statement has been summarized in the following table:

Table 1.1: Summary of Problem Statement

PS	Problem Statement
PS_1	Different possible rate of false alarm for unforeseen attack which remain as challenging task
PS_2	Unresolved issues such as predicting an intrusion as normal instances and normal instances as attacks or intrusion become inevitable limit in building effective anomaly detection

1.4. PROJECT QUESTION (PQ)

1. How to achieve the best possible rate of false alarm?
2. How to improve the false alarm by the hybridized method?
3. How to differentiate between the normal and non-normal attack correctly?

1.5. PROJECT OBJECTIVE (PO)

The ability to identify attempts to exploit new and unforeseen vulnerabilities is a noteworthy advantage for anomaly-based IDS. Data mining is the most recent technology introduced in network security environment to find regularities and irregulaties in large datasets.

Clustering is one of the anomaly detection methods which able to detect novel attack without any prior notice and is capable to find natural grouping of data based on similarities among the patterns. Clustering is a type of unsupervised learning. In this project, I investigate partition clustering using *K*-Means. *K*-Means clustering generally divide a dataset into several clusters where instances in the same cluster share certain properties together and are similar to each other to some extent. The main objective of implementation K-Means clustering is to split and group data into normal and attack instances while for classification is a type of supervised learning that is used to classify data into specific category as the classification is also one of the anomaly detection methods. Under classification approach, there are varieties of classifiers which have been widely cited. There is a requirement for an approach that could detect and identify