ANOMALY INTRUSION DETECTION SYSTEM USING IMMUNE NETWORK WITH REDUCED NETWORK TRAFFIC FEATURES

MURAD ABDO RASSAM QASEM

UNIVERSITI TEKNOLOGI MALAYSIA

ANOMALY INTRUSION DETECTION SYSTEM USING IMMUNE NETWORK WITH REDUCED NETWORK TRAFFIC FEATURES

MURAD ABDO RASSAM QASEM

A project report submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

APRIL 2010

To my beloved mother and father

To all my brothers and sisters

To my wife and children

To all my friends

To my beloved country, Yemen

ACKNOWLEDGEMENT

First and foremost, all praise and thanks are due to Allah, and peace and blessings be upon his Messenger, Mohammed (Peace Be Upon Him). Next, I wish to express my sincere appreciation to my main supervisor, Prof.Dr. Mohd Aizaini Maarof, for encouragement, guidance, critics, and friendship. I am also very thankful to my co-supervisor Mrs. Anazida Zainal for her patience and considerate nature that made her accessible whenever I needed her assistance. I indeed thank both of them for showing me how to identify interesting problems and how the research can be started and finished correctly.

I acknowledge that my UTM colleagues are the greatest. My especial thank to my brothers, Mohammed Al-shargabi, Qais Al-nuzaili, Ahmed Jabr, Adeeb Al-amery ,Waleed Al-odaini, and Den Fairol for their support and encouragement to get this work done. Their views and tips were useful indeed.

My sincere appreciation also extends to all my colleagues who have provided assistance at various occasions. I am grateful to all my family members. In particular, I would like to thank my wife for her patience, encouragement, support and understanding.

ABSTRACT

Intrusion Detection Systems (IDS) are developed to be the defense against these security threats. Current signature based IDS like firewalls and anti viruses, which rely on labeled training data, generally can not detect novel attacks. A method that offers a promise to solve this problem is the anomaly based IDS. Literature has shown that direction towards reducing false positive rate and thus enhancing the detection rate and speed have shifted from accurate machine learning classifiers to the adaptive models like bio-inspired models. Consequently, this study has been introduced to enhance the detection rate and speed up the detection process by reducing the network traffic features. Moreover, it aimed to investigate the implementation of the bio-inspired Immune Network approach for clustering different kinds of attacks. This approach aimed at enhancing the detection rate of novel attacks and thus decreasing the high false positive rate in IDS. Rough Set method was applied to reduce the dimension of KDD CUP '99 dataset which used by this study and select only the features that best represent all kinds of attacks. Immune Network clustering was then applied using aiNet algorithm in order to cluster normal data from attacks in the testing dataset. The results revealed that detection rate and speed were enhanced by using only the most significant features. Furthermore, it was found that Immune Network clustering method is robust in detecting novel attacks in the test dataset. The principal conclusion was that IDS is enhanced by the use of significant network traffic features besides the implementation of the Immune Network clustering to detect novel attacks.

ABSTRAK

Sistem Pengesanan Pencerobohan (IDS) dibangunkan untuk menangani ancaman keselamatan ini. Sistem pengesanan berteraskan tandatangan seperti dinding api dan anti-virus kebiasaannya tidak dapat mengecam serangan-serangan baru manakala sistem pengesanan berbentuk ‘anomaly’ berupaya menyelesaikan masalah sebegini. Kajian menunjukkan tumpuan telah beralih kepada pengurangan *false alarm* serta pembaikan terhadap tahap pengesanan dan kelajuan pengesanan dengan aplikasi seperti pengesanan berbentuk ‘machine-learning’ kepada kaedah berdasarkan ‘bio-inspired’. Oleh itu, kajian projek ini dibangunkan untuk menambahbaik kadar pengesanan serta kelajuan proses pengesanan ini dengan mengurangkan atribut pada paket rangkaian trafik. Secara spesifiknya, fokus kajian projek ini tertumpu kepada pengaplikasian pendekatan ‘bio-inspired Immune Network’ pada sistem pengesanan pencerobohan dengan mengumpukkan (cluster) kepada kelas-kelas serangan. Tujuannya adalah untuk menambahbaik kadar pengesanan terhadap serangan-serangan baru dan menurunkan kadar ‘false positive’. Kaedah ‘Rough Set’ digunakan untuk mengurangkan dimensi atribut pada paket-paket rangkaian set data KDD CUP ’99 serta memilih atribut-atribut yang terbaik bagi mewakili semua jenis serangan. Algoritma aiNet digunakan bagi mengasingkan kan data normal dari data serangan pada set data pengujian. Hasil kajian menunjukkan kadar pengesanan serta kelajuan proses pengesanan boleh dicapai dengan menggunakan atribut-atribut rangkaian yang penting. Disamping itu, Immune Network berupaya mengesan jenis-jenis serangan yang baru. Kesimpulannya, penambahbaikan pada sistem pengesanan pencerobohan ini boleh diperolehi dengan menggunakan atribut-atribut rangkaian yang penting sahaja mewakili semua jenis serangan dan Immune Network berupaya mengklusterkan serangan-serangan baru.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Due to the increase use of computer networks in many aspects of life, the number of vulnerabilities also is increasing causing the network resources unavailable and violates the system confidentiality, integrity and availability. Intrusions pose a serious security threat for the stability and the security of information in the network environment. A network intrusion attack encompasses a wide range of activities. It includes attempting to destabilize the network, gaining unauthorized access to files with privileges, or mishandling and misusing of software (Jiang *et al.*, 2006).

Intrusion Detection Systems (IDS's) are security tools that, like other measures such as antivirus software, firewalls, and access control schemes, are intended to strengthen the security of information and communication systems (Teodoro, 2009). An Intrusion Detection System is an important component of the computer and information security framework. Its main goal is to differentiate

between normal activities of the system and behaviors that can be classified as intrusive.

The purpose of IDS is to detect unauthorized use or access to the computer system or network from the outside environment by those who do not have the authority or access rights to such systems. The main goal of intrusion detection is to build a system that could automatically scan the network activity and detect such intrusion attacks. An IDS is used to detect several types of malicious behaviors that can compromise the security and trust of a computer system or network. This includes network attacks against vulnerable services, data driven attacks on applications, host based attacks such as privilege escalation, unauthorized logins and access to sensitive files, and malware (viruses, Trojan horses, and worms).

There are two main intrusion detection approaches: anomaly intrusion detection system and misuse intrusion detection system .The anomaly detection focuses on the unusal activities of patterns and uses the normal behavior patterns to identify an intrusion. The misuse detection recognizes known attack patterns and uses well-defined patterns of the attack. On the other hand, IDS's may be categorized according to the host system into two types:

i. Host-based IDS (HIDS)
ii. Network-based IDS(NIDS)

The first operates at the host level and monitors a single host machine using the audit trails of the host operating system, whereas the other operates at the network level and monitors any number of hosts on the network.

According to Zainal *et al*., (2006), IDS can be treated as pattern recognition problem or rather classified as learning system. They stated that, an appropriate representation space for learning by selecting relevant attributes to the problem

domain is an important issue for learning systems. Bello *et al., (*2005) suggested that feature reduction is necessary to reduce the dimensionality of training dataset. They claimed that feature reduction also enhances the speed of data manipulation and improves the classification rate by reducing the influence of noise.

Furthermore, Kim *et al.,* (2005) stated that the goal of feature reduction is to find a feature subset in order to maximize some performance criterion, such as classification accuracy. They claimed that selecting important features is an important issue in intrusion detection.

In literature, numbers of anomaly detection systems were developed based on many different machine learning techniques. For example, some studies apply single learning techniques, such as neural networks, genetic algorithms, support vector machines, bio-inspired algorithms , etc. On the other hand, some systems are based on combining different learning techniques, such as hybrid or ensemble techniques. In particular, these techniques were developed as classifiers, which were used to classify or recognize whether the incoming network access is normal access or an attack.

Computing models inspired by biology are a way to make use of concepts, principles and mechanisms underlying biological systems. This type of computing includes among others, fields as evolutionary algorithms, neural networks, molecular computing quantum computing and immunological computation. The trend now is going towards the bio-inspired systems because of the ability of those systems to adapt naturally with the environment in which they applied. The human immune system provides inspiration for solving a wide range of innovative problems.

The Immune Network theory as originally proposed by Jerne, (1974a) hypothesized a novel viewpoint of lymphocyte activities, natural antibody

production, pre-immune repertoire selection, tolerance and self/non-self discrimination, memory and the evolution of the immune system. It was suggested that the immune system is composed of a regulated network of cells and molecules that recognize one another even in the absence of antigens. The immune system was formally defined as an enormous and complex network of paratopes that recognize sets of idiotopes, and of idiotopes that are recognized by sets of paratopes, thus it could recognize as well as be recognized (de Castro *et al.,* 2001).

By dealing with very large number of data over networks, it is very difficult to classify them manually to detect possible intrusions. Labeled data can be obtained by simulating intrusions, but this is limited to the set of known attacks and will fail to address the new types of attacks that may occur in the future. As a result of that limited ability in detecting unknown attacks, the detection system will not be able to play its role in securing the network data. So a technique for detecting intrusions when the data is unlabeled is needed, as well as detecting new and unknown types of attacks .

1.2 Problem Background

Kim and Park, (2003) proposed a network-based Support Vector Machine SVM IDS, and demonstrated it through results of 3 kinds of experiments. They have shown that SVM IDS can be an effective choice of implementing IDS. They discovered that there are some miss-classified input vectors and those degrade the performance of SVM IDS. They claimed that the performance of SVM IDS can be improved by applying Genetic Algorithm (GA) based feature extraction. They further suggested that Decision Tree (DT) method can also be used to extract features instead of GA.

Shon *et al.*, (2005) employed a GA for selecting proper TCP/IP packet fields to be applied to support vector learning in order to distinguish anomaly attacks from normal packets. Their results showed that their proposed GA and the time delay preprocessing were reasonable for feature reduction. Moreover, they claimed that the two approaches using supervised and unsupervised SVM provide a high correction rate, but high false positive alarms.

A multi-level hybrid classification model combining DT and Bayesian Network (BN) clustering has been proposed by Xiang, (2008). Their results on KDD CUP'99 dataset, a benchmark dataset used to evaluate IDS, were compared with other popular approaches such as multi-level tree classifier and winners of KDD CUP'99. It was shown that this new approach is very efficient in detecting intrusions with an extremely low false-negative rate of 3.23%, while keeping an accepTable level of false-positive rate of 3.2%. Although the false positive rate was reported to be as low as 0.5% for the KDDCUP'99 winner (Pfahringer, 2000), the corresponding false negative rate was merely 9.1%, which is much higher than their result (3.23%).

Zanero and Savaresi, (2004) stated that the problem of IDS does not lie only in the sheer number of vulnerabilities that are discovered every day. They claimed that there are also an unknown number of unexposed vulnerabilities that may not be immediately available to the experts for analysis and inclusion in the knowledge base. In order to overcome this problem, they introduced an unsupervised anomaly detection based on clustering. They satated that their approach increase the detection rate of different kinds of unknown attacks

In most circumstances, labeled data or purely normal data is not readily available since it is time consuming and expensive to manually classify it. Purely normal data is also very hard to obtain in practice, since it is very hard to guarantee that there are non intrusions when they were collecting network traffic (Leung and

Leckie, 2005). They stated that to address these problems, they used a new type of intrusion detection algorithm called unsupervised anomaly detection.

Data reduction can be achieved by filtering, data clustering and feature selection (Chebrolu *et al*., 2004). Generally, the capability of anomaly intrusion detection is often hindered by the inability to accurately classify a variation of normal behavior as an intrusion. Additionally, network traffic data is huge, and it causes a prohibitively high overhead and often becomes a major problem in IDS (Sung and Mukkamala, 2004).

According to Chakraborty, (2005), the existence of these irrelevant and redundant features generally affects the performance of machine learning or pattern classification algorithms. Hassan, *et al*., (2003) proved that proper selection of feature set has resulted in better classification performance. (Sung and Mukkamala, 2004) have demonstrated that the elimination of these unimportant and irrelevant features did not significantly lowering the performance of IDS.

Chebrolu *et al*., (2004) tackled the issue of effectiveness of an IDS in terms of real-time and detection accuracy from the feature reduction perspective. In their work, features were reduced using two techniques, Bayesian Network (BN) and Classification and Regression Trees (CART). They have experimented using four sets of feature subset which are 12, 17, 19 and all the variables (41) from one network connection. Dataset used was KDD CUP 99. They suggested that using the full 41 features do drop the detection rate of the IDS.