COMPARATIVE STUDY BETWEEN FUZZY C-MEANS ALGORITHM AND ARTIFICIAL IMMUNE NETWORK ALGORITHM IN INTRUSION DETECTION SYSTEM

AHMED A. A. ABUNADA

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

Faculty of Computing
Universiti Teknologi Malaysia

JANUARY 2013

I dedicate this project to my respected and beloved my mother ‘Ibtesam Saed Al-khateb’ and My father ‘Abdurabou Ahmed Abunada’, thank you for the moral and financial support you have given me throughout my academic life.

To my respected supervisor, Dr. Anazida Zainal

To my beloved country, Palestine, Gaza

To all my brothers and sisters

To all my friends

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, **Dr. Mrs. Anazida Zainal,** for encouragement, guidance, critics, and friendship.

.

ABSTRACT

Intrusion Detection Systems (IDS) is special software developed in order to protect the system against security threats and malware. IDS provides second line of defense after rule based firewall. Unfortunately IDS with supervised learning approach heavily rely on labeled training data and generally it fails to detect novel attacks and produces high false alarm. Besides, data labeling is expensive and time consuming. However, a systematic method which offers the capability to alleviate this problem is through the use of unsupervised approach~~es~~, which is the basis for this research. In addition to that, to investigate this phenomenon, a comparison between two clustering algorithms based on an anomaly detection system IDS is proposed. Related literature has given a direction towards comparing two clustering algorithm which are Artificial Immune Network (AIN) and Fuzzy c-means (FCM). The performance of those two clustering algorithm were measured based on false positive rate, false negative rate, hit rate and detection. This study has evaluated and analyzed AIN and FCM clustering algorithms. The finding shows that AIN gives higher overall accuracy and hit rate. It also gives lower false alarms on both datasets used in the study. Consistent good performances of AIN in clustering network traffic data into respective classes has made AIN a promising clustering technique to be of used in detection novel attack traffic in IDS.

ABSTRAK

Sistem Pengesan Pencerobohan (IDS) adalah sebuah perisian khas yang dibangunkan untuk melindungi sistem dari ancaman keselamatan dan juga malware. IDS menyediakan pertahanan dua lapisan kepada firewall berasaskan peraturan. Malangnya IDS dengan pendekatan seliaan pembelajaran tersangat bergantung kepada data latihan berlabel dan umumnya ia gagal untuk mengesan serangan-serangan dan juga menghasilkan penggera palsu. Selain itu, pelabelan data adalah mahal dan memakan masa. Walaupun begitu, kaedah sistematik yang menawarkan keupayaan untuk mengatasi masalah ini adalah dengan menggunakan pendekatan tanpa pengawasan di mana adalah asas kepada penyelidikan ini. Tambahan pula, untuk menyiasat fenomena ini, perbandingan di antara dua kelompok algoritma berdasarkan sistem pengesanan anomali diusulkan. Penulisan berkaitan memberikan tunjuk arah terhadap perbandingan dua kelompok algoritma iaitu Artificial Immune Network (AIN) dan juga Fuzzy C-Means (FCM). Prestasi kedua-dua kelompok algoritma ini telah diukur berdasarkan nisbah positif palsu, nisbah negatif palsu, nisbah terkena ancaman dan juga pengesanan. Kajian ini telah menilai dan telah menganalisa algoritma kelompok AIN dan juga FCM. Hasil dari penemuan menunjukkan bahawa AIN mendapat ketepatan keseluruhan dan juga nisbah terkena ancaman yang tinggi. Ia juga mendapat penggera palsu yang rendah dalam kedua-dua set data yang digunakan dalam kajian ini. Prestasi baik AIN yang konsisten di dalam mengkelompokan data-data trafik rangkaian ke dalam kelas masing-masing telh membuat AIN teknik mengkelompokan yang memberangsangkan untuk kegunaan pengesanan trafik baru di dalam IDS.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

IDS	-	Intrusion detection system
AIN	-	Artificial Immune network
FCM	-	Fuzzy c-means
FP	-	False positive
FN	-	False Negative
HIDS	-	Host-based Intrusion detection system
NIDS	-	Network-based Intrusion detection system
ROC	-	Receiver Operating Characteristics
DoS	-	Denial of Service Attack
U2R	-	User to Root Attack
R2L	-	Remote to Local Attack
SOM	-	Self-Organizing Maps
aiNet		Artificial Immune Network

CHAPTER 1

INTRODUCTION

1.1 Introduction

Because of the growing utilization of computer networks in lots of facets of our existence, the amount of weaknesses is also growing leading to the network assets not available and split up the machine discretion, integrity and availability. Makes use of pose a significant security threat for that stability and also the security of knowledge within the network atmosphere. According to Qasem (2010), network invasion attack involves an array of activities. It offers trying to destabilize the network, attaining unauthorized use of files with rights, or mishandling and misusing of software.

Intrusion Detection Systems (IDS's) are security tools, like other measures for example anti-virus programs, firewalls, and access control schemes, usually are meant to strengthen the safety of knowledge and communication systems (Garcia-Teodoro *et al.*, 2009). An Intrusion Detection System is a vital element of the computer and knowledge security framework between normal activities from the system and actions that may be considered intrusive.

The objective of IDS would be to identify unauthorized use or accessibility computer or network in the outdoors atmosphere by individuals who do not possess the authority or access privileges to such systems. The primary purpose of Intrusion

Detection would be to develop a system that may instantly scan the network activity and identify such invasion attacks (Qasem, 2010). An IDS can be used to identify several kinds of malicious actions that may compromise the safety and trust of the computer or network. Including network attacks against vulnerable services, data driven attacks on programs, host based attacks for example privilege escalation, unauthorized logins and access to sensitive files, and malware.

There are two main intrusion detection approaches, misuse intrusion detection system and anomaly intrusion detection system exist. The Misuse intrusion detection is based on attack signatures, the detailed description of the sequence of actions performed by the attacker. This approach provides the platform which allows the detection of intrusions perfectly matching the signatures. On the other hand, the misuse detection recognizes known attack patterns and uses well-defined patterns of the attack.

The anomaly detection concentrates on the unusual activities of designs and uses the standard behavior designs to recognize an intrusion. Many researchers mentioned that the anomaly intrusion detection can solve the issues that misuse detection cannot solve. Garcia-Teodoro *et al*., (2009) highlighted that the primary advantage of anomaly detection approach is its ability to identify or detect previously unknown intrusions. Panda and Patra (2005), mentioned that, it's a must have way of discovering makes use of once the training information is unlabeled too for discovering unknown kinds of makes use of. They further stated that, the technique that satisfies this require is the anomaly detection and without supervision approach.

Meanwhile IDS's can also be categorized according to the host system into two types:

1. Host-based IDS (HIDS)
2. Network-based IDS(NIDS)

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

1.2 Problem Background

By dealing with very large number of data over networks, it is difficult to classify them manually to detect possible intrusions. Labeled data could be acquired by simulating intrusions, but this really is restricted to the group of known attacks and can neglect to address new kinds of attacks that could occur later on. Consequently of the limited ability in discovering unknown attacks, the recognition product is not efficient in acquiring the network data (Qasem, 2010). Therefore, a procedure for discovering makes use of once the information is unlabeled is required, in addition to discovering new and unknown kinds of attacks.

Anomaly detection algorithms hold the advantage that they may identify new types of intrusions as diversions from normal usage, Leon *et at*., (2004). Going by this problem and given some normal data to train from, and given a totally new bit of test data, invasion recognition formula is always to decide if test data take part in "normal" order to be able to detect an anomalous behavior. Referring to this issue as supervised anomaly detection because the models are produced only using the normal behavior across the network. In comparison, without supervision anomaly detection attempts to recognize anomalous behavior without needing any understanding regarding the training data. However, both kinds of anomaly detection schemes are stricken by maximum false alarms.

Qasem (2010) maintains that, in many conditions, labeled information is unavailable and the time is right consuming and incredibly costly to label the information by hand. Meanwhile, there's always an engaged alternation in normal traffic designs as well as constantly emerging of novel attacks each one of these

problems result in the supervised approach not practical solution for IDS. To resolve these complaints, scientists proceed to focus on without supervision approach, for example clustering because this without supervision approach doesn't rely on the labeled data and it doesn't consume us just as much time as supervised needs. Without a doubt, clustering will work for new novel attacks. Various without supervision techniques happen to be suggested however the recognition rate of IDS is quite insufficient compared to supervised approaches.

Bace and Mell (2001) suggest that, to manage to identify novel attacks, anomaly-based Intrusion detection systems was recommended. The job starting with modeling a range of normal or valid behavior, especially when the observed behavior diverges from this model, then an anomaly is elevated. However, anomaly-based IDSs are more likely to false positives that may be triggered by novel, but non-malicious traffic, as it is difficult to make a model connected wonderful possible normal traffic. These false positives generally are a considerable hindrance to effective operators monitoring the NIDS, consequently of occasions wasted in considering them. Single Percent false positive rate might trigger huge amounts of bogus alerts particularly when run on the large volumes of traffic common in current systems. This, according to Axelsson (2000), is known as the base rate fallacy. Nonetheless, anomaly-based approach has remained an active part of research interest and is the main focus of this research also.

Artificial immune system technique was introduced in late 90’s and it received a lot of attention from researchers. The ability of immune system to protect human body were adopted many algorithms such as Negative selection, Clonal selection and immune network. Applications Artificial Immune system include that of computer and internet security, network intrusion detection and computer viruses.

Immune network which is a clustering technique was founded by Jerne’s idiotypic network theory (Jerne1974), which suggests that the immune systems looks after a network of interconnected B-cells. In artificial immune network (AIN) models, a B-cell population includes two sub-populations: the very first population

as well as the cloned population. The very first set is created in the subset of raw training data to create the B-cell network.

1.3 Problem Statement

Supervised techniques do suffer low detection accuracy and high false alarm especially when dealing with novel attacks. Besides, labeling network traffic instances is expensive and time consuming. Furthermore, a supervised technique can be obsolete, especially when network traffic is dynamic. It warrants an updating of reference model. Therefore, clustering often seen to be a better solution as it can deal with changes.

1.4 Purpose of Study

In this research the performance of Fuzzy c-means algorithm (FCM) as well as Artificial Immune Network algorithm(AIN) will be compared in terms of detection accuracy , false alarms and hit rate. At the end of this comparison, an analysis of their performances will be discussed and the algorithm that shows better performance will be highlighted and recommended.

1.5 Objectives of Study

This research has the following objectives:

i. To study and investigate performance of Fuzzy c-Means algorithm in IDS .
ii. To study and investigate performance of Immune Network algorithm in IDS.

iii. To compare the performance of Fuzzy c-Means algorithm and Immune network algorithm in IDS.

1.6 Scope of Study

The scope of project is listed below:

i. Two clustering technique (Fuzzy c-means) and (Artificial Immune network) will be used in this study.
ii. The data used in this study is from KDD Cup 1999 Intrusion Detection dataset.
iii. The study intends to use two datasets which will comprise of 5,092 and 6,890 samples in order to retain actual distribution of KDD Cup 1999 data.
iv. Matlab will be used to code Fuzzy c-means (FCM), and Artificial Immune Network (AIN) algorithms.
v. The classification will be based on five classes which are Normal , Probe, DoS,U2R and R2L. as in works of (Abraham and Grosan, 2006, Zainal *et al.*, 2009), (Dutta, 2009)
vi. Performance will be evaluated based on detection accuracy , False positive rate , False negative rate and Hit rate.

1.7 Significant of Study

This study evaluates the performance of two algorithms: Artificial Immune Network clustering and Fuzzy c-means algorithm for the network-based IDS in terms of detection accuracy, and false alarms by studying each one and investigate them to show which one is more suitable to be used in IDS.

1.8 Organization of Report

The thesis consists of 6 chapters. Chapter one describes the introduction, background of the study, research objectives and questions, the scope of the study and its primary objectives. The second chapter reviews available and related literature on Intrusion detection Systems, Artificial Immune Network, supervised and without supervision, Fuzzy c-means and clustering approaches. Chapter three describes the study methodology along with the appropriate framework for the study. The 4th chapter provides the results and analysis of the findings of the first algorithm which is Fuzzy c-means (FCM). The 5th chapter provides the results and analysis of the findings of the second algorithm which is Artificial Immune Network (AIN) and the evaluation with Fuzzy c-means based on the detection Accuracy , False alarms and the Hit rate. Lastly, chapter 6 covers the conclusion and the future works

REFERENCES

Abraham, A.and Grosan, C. (2006). Evolving intrusion detection systems. *Genetic Systems Programming***,** 57-79.

Abraham, A., Grosan, C.and Ramos, V. (2006). *Swarm intelligence in data mining*, Springer-Verlag New York Inc.

Aickelin, U., Greensmith, J.and Twycross, J. (2004). Immune system approaches to intrusion detection–a review. *Artificial Immune Systems***,** 316-329.

Amasyali, F.and Albayrak, S. (2003). Fuzzy c-means clustering on medical diagnostic systems..

Budayan, C., Dikmen, I.and Birgonul, M. T. (2009). Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping. *Expert Systems with Applications,* 36**,** 11772-11781.

Chandola, V., Banerjee, A.and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR),* 41**,** 15.

Chimphlee, W., Abdullah, A. H., Sap, M. N. M., Chimphlee, S.and Srinoy, S. (2005). Unsupervised clustering methods for identifying rare events in anomaly detection. *a a,* 2**,** 1.

Davis, J. J.and Clark, A. J. (2011). Data preprocessing for anomaly based network intrusion detection: A review. *Computers & Security*.

Dutta, A. (2009). *Fuzzy c-Means Classification of Multispectral Data Incorporating Spatial Contextual Information by using Markov Random Field.* M. Sc Thesis, GFM, IIRS-ITC JEP.

Jiang, S. Y., Song, X., Wang, H., Han, J. J.and Li, Q. H. (2006). A clustering-based method for unsupervised intrusion detections. *Pattern Recognition Letters,* 27**,** 802-810.

Kiang, M. Y.and Fisher, D. M. (2008). Selecting the right MBA schools–An application of self-organizing map networks. *Expert Systems with Applications,* 35**,** 946-955.

Kim, J.and Bentley, P. J. (2001). An evaluation of negative selection in an artificial immune system for network intrusion detection.. San Francisco, California, USA, 1330-1337.

Laskov, P., Däussel, P., Schäafer, C.and And Rieck, K. (2005). Learnin intrusion detection:supervised or unsupervised?. Lecture Notes in Computer Science,

Lippmann, R. P., Fried, D. J., Graf, I., Haines, J. W., Kendall, K. R., Mcclung, D., Weber, D., Webster, S. E., Wyschogrod, D.and Cunningham, R. K. (2000). Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation.. IEEE, 12-26 vol. 2.

Mehrotra, K., Mohan, C. K.and Ranka, S. (1997). *Elements of artificial neural networks*, the MIT Press.

Nezakatolhoseini, M., Jabbehdari, S.and Pourmina, M. A. (2012). Analysis and Performance Evaluation of Application Specific Processors for Network-Based Intrusion Detection Systems. *Advances in Computing and Information Technology***,** 61-70.

Panda, M.and Patra, M. (2005). SOME Clustering Agorithm to enhance the performance of network intrusion detection system. Journal of Theoretical and Applied Information Technology.

Qasem, M. a. R. (2010). *Anomaly intrusion detection system using immune network with reduced network traffic features.* Universiti Teknologi Malaysia, Faculty of Computer Science and Information Systems.

Tavallaee, M., Bagheri, E., Lu, W.and Ghorbani, A. A. (2009). A detailed analysis of the kdd cup 99 data set.

Velmurugan, T.and Santhanam, T. (2010). Computational complexity between k-means and k-medoids clustering algorithms for normal and uniform distributions of data points. *Journal of Computer Science,* 6**,** 363-368.

Yasami, Y., Khorsandi, S., Mozaffari, S.and Jalalian, A. (Year). An unsupervised network anomaly detection approach by k-Means clustering & ID3 algorithms. *In*, 2008. IEEE, 398-403.

Zainal, A., Maarof, M. A.and Shamduddin, S. (2006). "Feature selection using rough set in intrusion detection", in Proc. IEEE TENCON, p.4.

Zainal, A., Maarof, M. A.and Shamsuddin, S. M. (2009). Ensemble classifiers for network intrusion detection system. *Journal of Information Assurance and Security,* 4**,** 217-225.