AN IMPROVED HYBRID LEARNING APPROACH FOR BETTER ANOMALY DETECTION

By

WARUSIA MOHAMED YASSIN

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirement for the Degree of Master of Science

May 2011

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Master of Science

AN IMPROVED HYBRID LEARNING APPROACH FOR ANOMALY DETECTION

By

WARUSIA MOHAMED YASSIN

May 2011

Chairman: Puan Hajah Zaiton Muda

Faculty: Computer Science and Information Technology

Intrusion Detection System (IDS) is facing complex requirements to overcome modern attack activities from damaging the computer systems. Gaining unauthorized access to files, attempting to damage the network and data, and any other serious security threat must be prevented by the Intrusion Detection System. Anomaly detection is one of intrusion detection techniques. This technique identifies an activity which deviates from the normal behaviours. Nonetheless, current anomaly detection techniques are unable to detect all types of attacks accurately and correctly. Therefore, anomaly detection is often associated with high false alarm with only moderate accuracy of detection rates.

In recent years, data mining approach for intrusion detection have been proposed and used such as neural networks, clustering, genetic algorithms, decision trees, and support vector machines. These approaches have resulted in high accuracy and good detection rates but with moderate false alarm on novel attacks. The recent works has been proposed by Tsai et

al. (2010) called a Triangle Area Based Nearest Neighbor (TANN) to obtain high accuracy and detection rate with low false alarms. Unfortunately this approach has not shown a remarkable improvement. In addition, some attacks and normal connections are even failed to be detected correctly. Therefore, there is a need for an approach that could detect and identify such attacks accurately in an interconnected network.

In this thesis, an improved hybrid mining approach is proposed through combination of *K*-Means clustering and classification techniques. *K*-Means clustering is an anomaly detection technique that is naturally capable for dealing with huge data in high speed network. *K*-Means clustering divides data into corresponding group called clusters, whereby all data in the same cluster are similar to each other. The proposed hybrid approach will be clustering all data into the corresponding group before applying a classifier for classification purposes. We choose k=3 in order to cluster data into three clusters called C1, C2 and C3. Probe, U2R and R2L attack data grouped into C1, while C2 is used to group DoS attack data. In order to separates normal data from an attack, C3 is used. Next, a number of classifiers like *Naïve Bayes*, *OneR*, and *Random Forest* separately applied to these data to group all data into the right categories.

An experiment is carried out to evaluate the performance of the proposed approach and the current techniques in terms of accuracy, detection rate, and false alarm rate using Knowledge Discovery in Databases (KDD) called KDD Cup ’99 intrusion detection dataset. The data covers four types of main attacks, which are *Denial-of-Services* (DoS), *User to Root* (U2R), *Remote to Local* (R2L), and *Probe*. Results show that the proposed approach performed better in term of accuracy, detection rates, and able to significantly reduce the false alarm rates.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains

PEREKAYASAAN PENDEKATAN PERLOMBONGAN PEMBELAJARAN BAGI PENGESANAN ANOMALI

Oleh

WARUSIA MOHAMED YASSIN

Mei 2011

Pengerusi: **Puan Hajah Zaiton Muda**

Fakulti: **Sains Komputer dan Teknologi Maklumat**

Sistem Pengesanan Pencerobohan (IDS) menghadapi cabaran yang kompleks dalam mengatasi aktiviti pencerobohan dan teknik serangan terkini daripada merosakkan sistem komputer. Perolehan capaian ke atas fail-fail, percubaan untuk merosakkan rangkaian data, serta lain-lain ancaman keselamatan yang serius perlulah dikesan oleh Sistem Pengesanan Pencerobohan. Teknik pengesanan anomali merupakan salah satu teknik pengesanan pencerobohan. Teknik ini mengenal pasti aktiviti-aktiviti yang tersasar daripada kelakuan normal. Walau bagaimanapun, teknik pengesanan anomali semasa tidak mampu mengesan kesemua jenis pencorobohan dengan betul dan tepat. Oleh itu, pengesanan anomali sering dikaitkan dengan amaran palsu yang tinggi serta ketepatan kadar pengesanan sederhana.

Kebelakangan ini, pendekatan perlombongan data seperti rangkaian neural, penggugusan, algoritma genetik, pepohon keputusan dan mesin vektor sokongan untuk pengesanan pencerobohan telah dicadangkan dan digunakan. Pendekatan ini memberikan hasil

ketepatan yang tinggi dan kadar pengesanan yang baik tetapi dengan amaran palsu yang agak sederhana ke atas serangan-serangan baharu. Tambahan itu, beberapa serangan dan penyambungan yang normal juga masih gagal dikesan dengan betul. Hasil kerja yang baru diperkenalkan oleh Tsai et al. (2010) iaitu “Triangle Area Based Nearest Neighbor” (TANN) untuk mengatasi masalah kadar pengesanan dan amaran palsu. Malangnya pendekatan ini tidak menunjukkan sebarang penigkatan. Oleh yang demikian, terdapat satu keperluan kepada pendekatan yang boleh mengesan dan mengenal pasti serangan-serangan secara tepat dalam sesebuah jaringan rangkaian.

Dalam tesis ini, sebuah pendekatan perlombongan hibrid yang direkayasa telah dicadangkan melalui penggabungan teknik penggugusan *K*-Means dan pengklasifikasian. Penggugusan *K*-Means adalah sejenis pengesanan anomali yang secara semulajadinya berupaya menguruskan kumpulan data yang banyak dalam rangkaian yang berkelajuan tinggi. Penggugusan *K*-Means membahagikan data ke dalam beberapa kumpulan yang dipanggil kelompok, yang mana data di dalam sesebuah kelompok mempunyai ciri yang sama antara satu dengan lain. Pendekatan hibrid yang dicadangkan akan mengasingkan kesemua data mengikut kelompok sebelum mengaplikasikan sebuah pengelas bagi tujuan klasifikasi. Oleh yang demikian, kami memilih k=3 untuk mengumpul kesemua data kedalam tiga kelompok iaitu C1, C2 dan C3. C1 digunakan untuk mengumpul data *User to Root* (U2R), *Remote to Local* (R2L) dan *Probe*, manakala C2 dugunakan untuk mengumpul data *Denial-of-Services* (DoS). Untuk memisahkan data *Normal* dari data-data serangan, C3 digunakan. Seterusnya, pengelas seperti *Naive Bayes*, *OneR* dan *Random Forest* digunakan secara berasingan ke atas data tersebut bagi pengkelasan ke dalam kategori yang betul.

Sebuah eksperimen telah dijalankan bagi menilai prestasi kaedah yang dicadangkan berbanding dengan teknik sedia ada dari segi ketepatan, kadar pengesanan dan kadar amaran palsu menggunakan set data “Knowledge Discovery in Databases (KDD)” KDD Cup '99. Data-data ini terbahagi kepada empat kelas serangan utama iaitu *Denial-of-Services* (DoS), *User to Root* (U2R), *Remote to Local* (R2L) dan *Probe*. Keputusan ekserimen menunjukkan bahawa pendekatan yang dicadangkan memberikan peningkatan prestasi dari segi ketepatan, kadar pengesanan dan keupayaan mengurangkan kadar amaran palsu ke tahap yang lebih rendah.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation and deepest gratitude to my supervisor Ms. Zaiton Muda and to both my committee members Prof. Madya Dr. Md. Nasir and Dr. Nur Izura Udzir for their continuous encouragement, valuable advices, and guidance throughout this research. I really appreciate the freedom they provide while I was working on my research and their openness to new ideas.

Special thanks to my dearest friends who are always willing to help and to share ideas and knowledge at times when they are busy with their own project themselves. I will treasure their friendship.

Most of all, I would like to express my sweetest appreciation to my family for their affectionate support, patience, and encouragement. Their prayers and good wishes always help me to be strong especially in difficult times. I am very grateful and thankful to them.

APPROVAL

I certify that an Examination Committee has met on _________________ to conduct the final examination of Warusia Mohamed Yassin on his Masters of Science thesis entitled "An Improved Hybrid Mining Approach for Anomaly Detection" in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The committee recommends that the student be awarded the degree of Master of Science.

Members of the Examination Committee are as follows:

Dr. Rodziah Atan, Ph.D.
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Dr. Norwati Mustapha, Ph.D.
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Dr. Zuriati Ahmad Zulkarnain, Ph.D.
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Dr. Rabiah Ahmad, Ph.D.
Associate Professor
Faculty of Information Technology and Communication
Universiti Teknikal Malaysia Melaka
(External Examiner)

____

BUJANG BIN KIM HUAT, Ph.D.
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Master of Science. The members of the Supervisory Committee were as follows:

Hajah Zaiton Muda
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Md. Nasir Sulaiman, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Nur Izura Udzir, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

____

HASANAH MOHD GHAZALI, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at Universiti Putra Malaysia or at any other institutions.

WARUSIA MOHAMED YASSIN

Date: 26 May 2011

TABLE OF CONTENTS

			Page
DEDICATION			ii
ABSTRACT			iii
ABSTRAK			vi
ACKNOWLEDGEMENTS			ix
APPROVAL			x
DECLARATION			xiiii
LIST OF TABLES			xvv
LIST OF FIGURES			xviiii
LIST OF ABBREVIATIONS			xixx
CHAPTER			
1	**INTRODUCTION**		**1**
	1.1	Background	1
	1.2	Problem statement	4
	1.3	Objectives of research	5
	1.4	Scope of research	6
	1.5	Organization of this thesis	6
2	**LITERATURE REVIEWS**		**8**
	2.1	Introduction8	
	2.2	Data mining techniques in anomaly detection	11
		2.2.1 Clustering methods	12
		2.2.2 Classification methods	13
	2.3	Hybrid machine learning	16
	2.4	Related works	17
	2.5	Summary	22
3	**RESEARCH METHODOLOGY24**		
	3.1	Introduction	24
	3.2	Research overview	24
		3.2.1 Stage 1: Data Preparation	25
		3.2.2 Stage 2: Clustering and Classification	26
	3.3	Research steps	26
		3.3.1 Problem identification	27
		3.3.2 Dataset preparation	27
		3.3.3 Design of the proposed method	28
		3.3.4 Implementation	28
		3.3.5 Experiment	29
		3.3.6 Analysis	30
	3.4	Summary	30

4 PROPOSED HYBRID MINING APPROACH **31**
4.1 Introduction 31
4.2 Proposed method 32
4.3 Naïve Bayes 36
4.4 Random Forest 38
4.5 OneR 40
4.6 Summary 44

5 RESULTS AND DISCUSSION **45**
5.1 Introduction 45
5.2 KDD Cup 99 dataset 45
5.3 Evaluation measurement 48
5.4 Performance evaluation 49
5.4.1 NB and (KM+NB) 50
5.4.2 RF and (KM+RF) 52
5.4.3 OR and (KM+OR) 54
5.4.4 Overall analysis and discussion 57
5.5 Performance validation 59
5.5.1 NB and (KM+NB) 59
5.5.2 RF and (KM+RF) 62
5.5.3 OR and (KM+OR) 64
5.5.4 Overall analysis and discussion 66
5.6 Performance comparison 69
5.7 Further comparison 72
5.8 Summary 74

6 CONCLUSION AND FUTURE WORK **75**
6.1 Conclusion 75
6.2 Future works 77

REFERENCES **79**
BIODATA OF STUDENT **86**

LIST OF TABLES

Page

Table 2.1: Comparisons of related work 20
Table 5.1: Sample distribution of the training dataset 47
Table 5.2: Sample distribution of the testing dataset 47
Table 5.3: KDD Cup '99 dataset attack type and pattern classification 48
Table 5.4: General behavior of intrusion detection data 49
Table 5.5: Classification result for each category class (NB) 50
Table 5.6: Classification result for each category class (KM+NB) 51
Table 5.7: Classification result for the normal and attack class (NB) 51
Table 5.8: Classification result for the normal and attack class (KM+NB) 51
Table 5.9: Classification result for each category class (RF) 52
Table 5.10: Classification result for the normal and attack class (KM+RF) 53
Table 5.11: Classification result for the normal and attack class (RF) 53
Table 5.12: Classification result for the normal and attack class (KM+RF) 53
Table 5.13: Classification result for each category class (OR) 55
Table 5.14: Classification result for each category class (KM+OR) 55
Table 5.15: Classification result for the normal and attack class (OR) 56
Table 5.16: Classification result for the normal and attack class (KM+OR) 56
Table 5.17: Single classifiers vs. hybrid approach using training dataset 57
Table 5.18: Classification result for each category class (NB) 60
Table 5.19: Classification result for each category class (KM+NB) 60
Table 5.20: Classification result for the normal and attack class (NB) 60
Table 5.21: Classification result for the normal and attack class (KM+NB) 61
Table 5.22: Classification result for each category class (RF) 62
Table 5.23: Classification result for each category class (KM+RF) 62
Table 5.24: Classification result for the normal and attack class (RF) 63
Table 5.25: Classification result for the normal and attack class (KM+RF) 63
Table 5.26: Classification result for each category class (OR) 64
Table 5.27: Classification result for each category class (KM+OR) 65
Table 5.28: Classification result for the normal and attack class (OR) 65

Table 5.29: Classification result for the normal and attack class (KM+OR) 65
Table 5.30: Single classifiers vs. hybrid approaches using testing dataset 67
Table 5.31: Previous vs. proposed hybrid approach using training dataset 69
Table 5.32: Previous vs. proposed hybrid approach using testing dataset 71
Table 5.33: Further comparison with previous findings 73

LIST OF FIGURES

	Page
Figure 1.1: Statistic of reported incident in 2010	2
Figure 1.2: Graph of reported incident in 2010	2
Figure 2.1: Network IDS architecture	9
Figure 3.1: Experimental and Implementation Process Flow	25
Figure 3.2: Research Steps	27
Figure 4.1: (a) Seeds	32
Figure 4.1: (b) Assigns instances to cluster	32
Figure 4.1: (c) Finds new centroid	33
Figure 4.1: (d) New	33
Figure 4.2: The K-Means algorithm	35
Figure 4.3: Classifier	36
Figure 4.4: Hybrid Algorithm for K-Means and Naïve Bayes	38
Figure 4.5: Random Forest algorithms	39
Figure 4.6: Hybrid Algorithm for K-Means and Random Forest	40
Figure 4.7: (a) Original data	41
Figure 4.7: (b) Step 1	42
Figure 4.7: (c) Step 2	42
Figure 4.7: (d) Step 3	43
Figure 4.7: (e) Result	43
Figure 4.8: Hybrid Algorithm for K-Means and OneR	44
Figure 5.1: Accuracy - Category classes (NB vs. KM+NB)	52
Figure 5.2: Accuracy - Category classes (RF vs. KM+RF)	54
Figure 5.3: Accuracy - Category classes (OR vs. KM+OR)	56
Figure 5.4: Accuracy	58
Figure 5.5: Detection Rate	58
Figure 5.6: False Alarm	58
Figure 5.7: Accuracy - Category classes (NB vs. KM+NB)	61
Figure 5.8: Accuracy - Category classes (RF vs. KM+RF)	64
Figure 5.9: Accuracy - Category classes (OR vs. KM+OR)	66

Figure 5.10: Accuracy 67
Figure 5.11: Detection Rate 67
Figure 5.12: False Alarm 68
Figure 5.13: Accuracy 70
Figure 5.14: Detection rate 70
Figure 5.15: False Alarm 70
Figure 5.16: Accuracy 71
Figure 5.17: Detection rate 72
Figure 5.18: False alarm 72

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
DoS	Derial-of-Services
DR	Detection Rate
FA	False Alarm
FN	False Negative
FNT	Flexible Neural Tree Model
FP	False Positive
H-SOM	Hierarchical Self-Organizing Maps
IDS	Intrusion Detection System
KDD	Knowledge Data Database
K-NN	*K*-Nearest Neighbor
My CERT	Malaysia Computer Emergency Response Team
NB	Naïve Bayes
OR	OneR
R2L	Root to Local
RF	Random Forest
KM+NB	*K*-Means+NaiveBayes
KM+OR	*K*-Means+OneR
KM+RF	*K*-Means+Random Forest
SOM	Self-Organizing Maps
SQL	Structure Query Language
SVM	Support Vector Machines
TANN	Triangle Area Nearest Neighbor
TN	True Negative
TP	True Positive
U2R	User to Root

CHAPTER 1

INTRODUCTION

1.1 Background

In present days, information security has become one of the important keys in our daily life. Little do we realize that computer users who are either connected through physical networks or in wireless environment are unaware of the fact that they are vulnerable to the risk of threats. Along with continuous expansion and growth in high-speed development of the Internet, sensitive and valuable information are scattered almost everywhere in the network, thus making the network environment to become complex than before. Although Internet provides real-time services and convenience to the users, there are issues in security of information, in which some bared to invasion threat. To date, servers are continuously being attacked and paralyzed, which costs huge monetary loss as well as business availability.

On 7th February 2000, Yahoo! suffered from DDOS attack and was paralyzed for three hours, affecting an approximate of one million users (Levine et al., 2000). One day after the incident, few other online service providers such as Amazon, Buy.com, CNN, and eBay also suffered from the same attacks with a combined calculated loss close to USD1.1 million. Figure 1.1 is a statistic from the Malaysia Computer Emergency Response Team (MyCERT) sourced from http://www.mycert.org.my, showing an increase number of attack reports growth on monthly basis throughout the year 2011.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	TOTAL
Content Related	0	2	4	1	4	3	5	4	10	1	2	3	39
Cyber Harrassment	11	21	25	20	22	20	36	32	61	54	63	54	419
Denial of Service	11	5	2	0	1	2	3	4	1	27	8	2	66
Fraud	159	111	176	176	137	111	153	167	181	298	340	203	2212
Intrusion	130	189	185	228	204	149	137	182	228	150	118	260	2160
Intrusion Attempt	17	12	38	46	51	49	108	77	113	32	75	67	685
Malicious Codes	50	66	104	91	108	78	102	141	113	131	110	105	1199
Spam	0	0	0	73	39	42	111	120	81	328	261	213	1268
Vulnerabilities Report	3	4	4	5	1	1	2	15	3	2	1	1	42
TOTAL	381	410	538	640	567	455	657	742	791	1023	978	908	8090

Figure 1.1: Statistic of reported incident in 2010

Figure 1.2 shows the statistic of reported incident based on general incident classification in 2010, sourced from http://www.mycert.org.my.

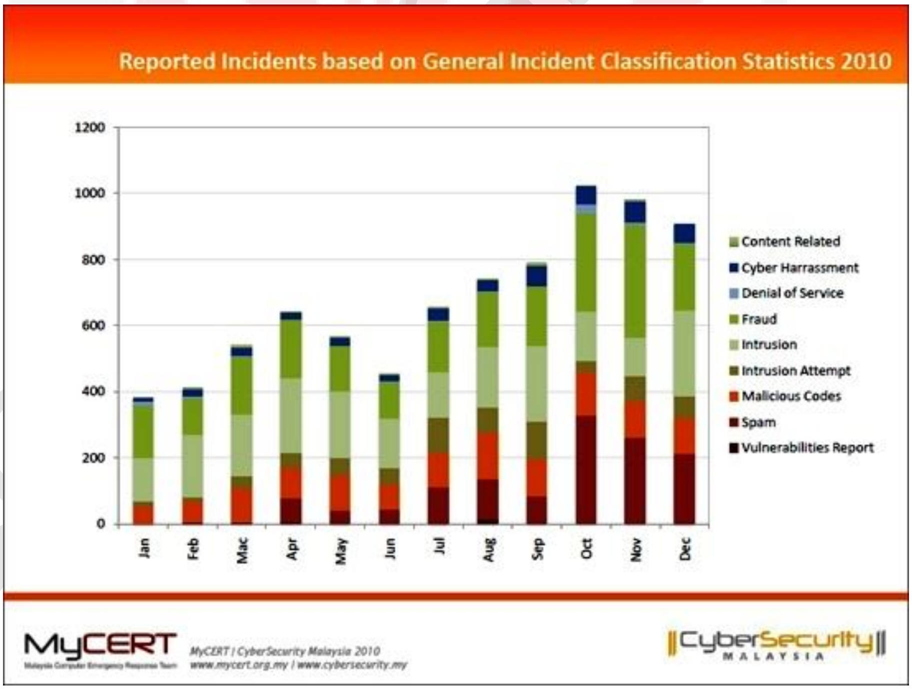

Figure 1.2: Graph of reported incident in 2010

Internet-based attacks have also become a new weapon in war. Back in the 1st September 2010, Indonesian hackers were reported to be drawing plan for mass defacement on

Malaysia (http://security.org.my/). The fact is that attackers are able to easily adapt and exploit new attack strategies without restriction with the help from Internet facilities at their convenience. 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 in security that suit the 'defense in depth' model is called the Intrusion Detection System (IDS) (Stephen et al., 2008). An IDS is capable of sending early alarm upon risk exposure caused by any attack. This is to alert the system administrators to execute corresponding response measurements, thus to reduce the possibility of bigger losses.

A growing interest in investigation of anomaly detection sparks from the ability of the approach to detect unknown attacks and to evaluate unforeseen vulnerability. Nonetheless, current anomaly detection technique suffers from high false alarm rate. Similarly, machine learning, being one of the most promising advancements in solving intricate data classification problems with accuracy also suffers from the same drawback. In view of this, this research proposes a new hybrid mining approach to improve current anomaly detection capabilities in IDS that would be an essential component of a security arsenal to fit the 'defense in depth' architecture in securing an information infrastructure.

1.2 Problem statement

The ultimate goal of anomaly detection in the development of IDS is to achieve the best possible accuracy and detection rate, as well as to reduce the rate of false alarm for every task at hand. Recently, there has been rigorous effort in improving the existing anomaly detection techniques due to significantly high false alarm as well as moderate accuracy and detection rate. In addition, there is lacking in performance of single classifier, which has resulted in high tendency for wrong classification during detecting unknown attacks (Tsai et al., 2010 and John et al., 2000). 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.

In short, a number of hybrid techniques have been proposed in intrusion detection fields which has been successfully identifying several novel intrusions correctly such as feature selection with SVM (Amiri et al., 2011), BIRCH Clustering with SVM (Horng et al., 2011), Triangle Area Based Nearest Neighbor (Tsai et al., 2010), ANN with Fuzzy Clustering (Gang et al., 2010), AIN with NN (Cao et al., 2010), Decision Tree with SVM (Su-Yun et al., 2009), Genetic Algorithm with SVM (Shon et al., 2007) and SOM with ANN (Liu et al., 2007); but there are still room to improve the accuracy and detection rate as well as the false alarm rate.

A potential drawback of all proposed approaches is the rate of false alarms with moderate accuracy and detection rate. To overcome these drawbacks, we proposed a combination of

K-Means clustering and classification techniques for intrusion detection which based on hybrid learning approach.

1.3 Objectives of research

The main objective of this research is to increase the accuracy and detection rate at lower false alarm rates by proposing an improved method. The proposed hybrid method is a combination of *K*-Means clustering and classification techniques. The *K*-Means clustering are required to cluster each and every data according to their group behavior. Next, the classifier techniques are applied to these clusters in order to classify the data into five categories including *U2R*, *R2L*, *Probe*, *DoS* and *Normal*.

1.4 Scope of research

We scope of this research to hybrid mining approach, which are use to analyze and find patterns in order to seperate an intrusion and normal instances correclty. There are two types of techniques chosen in this research, which is *K*-Means clustering and classification. The hybrid algorithms will be tested against KDD Cup ’99 dataset, a common benchmark intrusion detection dataset used to evaluate intrusion detection techniques. KDD Cup ‘99 has various attacks presented in the testing and validation data, making task more realistic and challenging in order to assess and validate the proposed approach based on percentage of accuracy, detection, and false alarm rate.

1.5 Organization of this thesis

This thesis is organized in accordance with the standard structure of thesis dissertations for Universiti Putra Malaysia. The thesis is inherently divided into five chapters as follows:

Chapter 1 – Introduction. This chapter introduces the degree of importance on information security and the general impact as globally. Awareness on current security issues forms the problem statement and the research objective.

Chapter 2 – Literature Reviews. This chapter reviews related studies on the fundamental knowledge of the subject matter such as Intrusion Detection Systems, signature-based detection, anomaly-based detection, hybrid learning techniques, data mining and other related techniques.

Chapter 3 – Research Methodology. This chapter presents an overview of research steps, which comprise of problem identification, dataset preparation, design of the proposed method, implementation of proposed method, and finally experiment and analysis.

Chapter 4 – Proposed Hybrid Mining Approach. This chapter introduces a combination of hybrid mining approach by applying classifier and *K*-Means clustering to anomalous instances using anomaly detection method. The approach is proposed to enhance the performance of overall single classifier in term of accuracy, detection and false alarm rate.

Chapter 5 – Result and Discussion. This chapter discusses the data source, medium of performance evaluation, as well as the experimental process flow that are adopted during the experiments. In addition, this chapter also provides a comparison between the proposed and the existing approaches.

Chapter 6 – Conclusion and Future Work. This chapter concludes the research with some recommendations for future work and development.

REFERENCES

Abadeh, M.S., Habibi, J., Barzegar, Z., and Sergi, M. (2007). A parallel genetic local search algorithm for intrusion detection in computer networks. Engineering Applications of Artificial Intelligence, 20:1058–1069.

Amiri, F., Mohammad, R. Y., Caro, L., Azadeh, S., and Nasser, Y. (2011). Mutual Information-Based Feature Selection for Intrusion Detection System. Journal of Network and Computer applications, 34: 1184–1199.

Anderson J.P. (1980). Computer security threat monitoring and surveillance. Techical Report.

Animesh Patcha and Jung-Min Park. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks. 51:3448–3470.

Arman, T., Mohammad, R., and Abdolreza, M. (2009). Intrusion detection using fuzzy association rules. Applied Soft Computing. 9 : 462–469.

Bace, R. and Mell, O. (2001). Intrusion detection systems. NIST Special Publications SP:800-31.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2):123–140.

Breiman, L. (1999). Random forests. Technical Report, Statistics Department, University of California, Berkeley. Available at http://www.stat.berkely.edu.

Breiman, L. (2001). Random forests. Machine Learning, 45:5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Books/Cole Advanced Boks & Software.

Brugger, S.T. (2004). Data mining methods for network intrusion detection. ISCR Computer Science Series, University of California.

Amiri, F., Mohammad, R. Y., Caro, L., Azadeh, S., and Nasser, Y. (2011). Mutual Information-Based Feature Selection for Intrusion Detection System. Journal of Network and Computer applications, 34: 1184–1199.

Chen, Y., Abraham, A., and Yang, B. (2007). Hybrid flexible neural-tree-based intrusion detection systems. International Journal of Intelligent Systems, 22:337–352.

Chih-Fong Tsai, Yu-Feng Hsu, Chia-Ying Lin and Wei-Yang Lin. (2009). Intrusion detection by machine learning: A review. Expert Systems with Applications. 36:11994–12000.

Denning D. (1987). An intrusion detection model. IEEE Transaction on Software Engineering, 13(2):222-232.

Endorf, C., Schultz, E., and Mellander, J. (2004). Intrusion detection and prevention. McGraw-Hill/Osborne.

Eric, B. (2001). Data Mining for Network Intrusion Detection: How to Get Started. Technical paper, The MITRE Corporation, Mclean, VA.

Eskin, E., Arnold, A., Preraua, M., Portnoy, L., and Stolfo, S.J. (2002). A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data, In D. Barbar and S. Jajodia (Eds.), Data Mining for Security Applications. Boston: Kluwer Academic Publishers.

Feng, H.H., Kolesnikov, O.M., Fogla, P., Lee, W., and Gong, W. (2003). Anomaly detection using call stack information. In Proceedings of the IEEE Symposium on Security and Privacy, Berkeley, CA, pp. 62-76.

Gandhi, M. And Srivatsa, S.K. (2008). Detecting and preventing attacks using network Intrusion detection systems. International Journal of Computer Science and Security, 2(1):49-60.

Gang, W., Jinxing, H., and Jian, M. (2011). A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering. Expert systems with applications,376: 6225–6232.

George H. John. (1995). Estimating Continuous Distributions in Bayesian Classifiers. In Proceedings of the Eleventh Conference on Uncertainly in Artificial Intellegence, Morgan Kaufmann Publishers, San Mateo.

Giancinto, G., Roli, F., and Didaci, L. (2003). Fusion of multiple classifiers for intrusion detection in computer network. Pattern Recognition Letter, ACM. 24(12): 1795-1803.

Guan, Y., A. Ghorbani and N. Belacel, “Y-means: A Clustering Method for Intrusion Detection,” Proceedings of Canadian Conference on Electrical and Computer Engineering. Montreal, Quebec, Canada, May, 2003.

Guba, S., Rastogi, R., and Shim, K. (1998). CURE: An efficient clustering algorithm for large databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM Press, pp. 73-84.

Harry, Z., and Jiang, S. (2008). Naive Bayes for optimal ranking. Journal of Experimental and Theoretical Artificial Intelligence. 20: 79-93.

Holt, R.C. (1993). Very simple classification rules perform well on most commonly used datasets. Machine Learning, 11: 69-90.

Horng, S.J. (2011). A novel intrusion detection system based on hierarchical clustering and support vector machines. Expert Systems with Applications. 38(1): 306-313.

Huy Anh, N., and Deokjai, C. (2008). Application of Data Mining to Network Intrusion Detection: Classifier Selection Model. Lecture Notes in Computer Science. 5297:399-408.

Jain, A. and Dubes, R. (1988). Algorithms for clustering data. Prentice-Hall, Englewood Cliffs, New Jersey.

Jang, J.S., Sun, C.T., and Mizutani, E. (1996). Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence. Prentice-Hall, Englewood Cliffs, New Jersey.

John McHugh, A.C., and Julia, A. (2000). Defending Yourself: The Role of Intrusion Detection Systems. Journal IEEE Software. pp. 42-51.

Kavitha, B., Karthikeyan, S., and Chitra, B. (2010). Efficient Intrusion Detection with Reduced Dimension Using Data Mining Classification Methods and Their Performance Comparison. Communications in Computer and Information Science. 1(70):96-101.

Kaufman, L. and Rousseeuw, P. (1990). Finding groups in data: An introduction to cluster analysis. John Wiley and Sons, New York, NY.

Kayacik, H.G., Nur, Z.H., and Heywood, M.I. (2007). A hierarchical SOM-based intrusion detection system. Engineering Applications of Artificial Intelligence, 20:439–451.

KDD. (1999). Available at http://kdd.ics.uci.edu/databases/-kddcup99/kddcup99.html

Khan, L., Awad, M., and Thuraisingham, B. (2007). A new intrusion detection system using support vector machines and hierarchical clustering. The VLDB Journal, 16:507–521.

Krista Rizman Zalik. (2008). An efficient *K*-Means clustering algorithm. Pattern Recognition Letter. 29: 1385-1391.

Latifur, K., Mamoun, A., and Bhavani, T. (2007). A New Intrusion Detection System Using Support Vector Machines And Hierarchical Clustering. The VLDB Journal. 16: 507–521.

Lee, W. and Stolfo, S. (2000). A framework for constructing features and models for intrusion detection systems. ACM Transaction of Information System Security 3(4):227–261.

Leung, K. And Leckie, C. (2005). Unsupervised anomaly detection in network intrusion detection using clusters, In Proceedings of the Twenty-eighth Australasian Conference on Computer Science, Newcastle, Australia, 38:333–342.

Levine, D.E. and Kessler, G.C. (2000). Defense against distributed denial of service attack. 4th edition of the Computer Security Handbook.

Li, Y. and Guo, L. (2007). An active learning based on TCM-KNN algorithm for supervised network intrusion. Computer and Securtiy, 26: 459-467.

Liang, H., Wei-Wu, R., and Fei, R. (2009). Anomaly detection using improved hierarchy clustering. In Proceedings of the International Conference on Artificial Intelligence and Computational Intelligence.

Lippmann, R.P., Fried, D.J., Graf, I., Haines, J.W., Kendall, K.R., McClung, D., Weber, D., Webster, S.E., Wyschogrod, D., Cunningham, R.K., and Zissman, M.A. (2000). Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. In Proceedings of the 2000 DARPA Information Survivability Conference and Exposition (DISCEX), Los Alamitos, CA, 2:12–26.

Liu, G., Yi, Z., and Yang, S. (2007). A hierarchical intrusion detection model based on the PCA neural networks. Neurocomputing, 70:1561–1568.

Liu, Y., Chen, K., Liao, X., and Zhang, W. (2004). A genetic clustering method for intrusion detection. Pattern Recognition, 37: 927–942.

Lodin S. (1998). Intrusion detection product evaluation criteria. Ernst and Young LLP, Available at http://docshow.net/ids.htm.

Lu J.W., Plataniotis, K.N., Venetsanopoulos, A.N., and Li, S.Z. (2006). Ensemble-based discriminant learning with boosting for face recognition. IEEE Transaction on Neural Networks, 17:166-178.

Luigi, R., Anderson, T.E., and McKeown, N. (2006). Traffic classification using clustering algorithms. In Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Sept. 11-15, Pisa, Italy, ACM Press, pp. 281-286.

Meera, G., and Srivatsa, S.K. (2010). Classification Algorithms in Comparing Classifier Categories to Predict the Accuracy of the Network Intrusion Detection – A Machine Learning Approach. Advances in Computational Sciences and Technology. 3(3):321–334.

Mrutyunjaya, P., and Manas, R.P.(2007). Network Intrusion Detection Using Naïve Bayes. International Journal of Computer Science and Network Security. 7(12):258-263.

Mihael, A., Markus, M, Breuig, H., Kriegel, P., and Sander, J. (1999). OPTICS: Ordering pionts to identify the clustering structure. In Proceedings of the ACM SiGMOD'99 International Conference on Management of data, ACM Press, pp. 49-60.

Ming, X., and Changjun, Z. (2009). Applied Research on Data Mining Algorithm in Network Intrusion Detection. International Joint Conference on Artificial Intelligence.

Nahla, B.A., Salem, B., and Zied, E. (2004). Naive Bayes vs Decision Trees in Intrusion Detection Systems. In Proceeding of the ACM Symposium on Applied Computing, Nicosia, Cyprus.

Ozyer, T., Alhajj, R., and Barker, K. (2007). Intrusion detection by integrating boosting genetic fuzzy classifier and data mining criteria for rule pre-screening. Journal of Network and Computer Applications, 30:99–113.

Panda, M. and Patra, M.R. (2008). A comparative study of data mining algorithms for network intrusion detection. In Proceedings of ICETET, India, pp.504-507.

Peddabachigari, S., Abraham, A., Grosan, C., and Thomas, J. (2007). Modeling intrusion detection system using hybrid intelligent systems. Journal of Network and Computer Applications, 30:114–132.

Pfahringer, B. (1999). Winning entry of the KDD '99 classifier learning contest. Available at http://www.acm.org/sigs/sigkdd/kddcup/1999.

Popescu, B.E. (2004). Ensemble learning for prediction. Stanford, CA, USA Stanford University.

Powers, S.T. and He, J. (2008). A hybrid artificial immune system and self-organising map for network intrusion detection. Journal of Information Sciences, 178: 3024-3042.

Rebecca Base and Peter Mell. (2001). NIST Special Publication on Intrusion Detection Systems. Infidel, Inc.,Scotts Valley, CA and National Institute of Standards and Technology.

Robert C. Holte. (1993). Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Machine Learning. 11: 63-91.

Rumelhart, D., Hinton, G., and Williams, R.. (1986). Learning internal representations by back-propagating errors. In D. Rumelhart and J. McClelland, editors, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, MIT Press, Cambridge, 1:318-362.

Russell, S.J. and Norvig, P. (2002). Artificial intelligence: A modern approach. Pearson US Imports & PHIPEs.

Sekar, R., Bendre, M., Dhurjati, P., and Bullineni, D. (2001). A fast automaton-based method for detecting anomalous program behaviours. In Proceedings of the IEEE Symposium on Security and Privacy, pp. 144–155.

Shaohua, T., Hongle, D., Naiqi, W., Wei, Z., and Jiangyi, S. (2010). A Cooperative Network Intrusion Detection Based on Fuzzy SVMs. Journal of Network s, 5: 475–483.

Shon, T. And Moon, J. (2007). A hybrid machine learning approach to network anomaly detection. Information Sciences, 177:3799–3821.

Shon, T., Kovah, X., and Moon, J. (2006). Applying genetic algorithm for classifying anomalous TCP/IP packets. Neurocomputing, 69:2429–2433.

Solahuddin, M. (2008). Applying knowledge discovery in database techniques. Modeling Packet Header Anomaly Intrusion Detection Systems, Journal of Software, 3(9): 68-76.

Stallings, W. (2006). Cryptography and network security principles and practices. Prentice-Hall, USA.

Steinberg, D., Golovnya, M., and Cardell, N.S. (2004). A brief overview to random forests. Available at http://www.salford-systems.com dstein@salford-systems.com.

Toosi, M. (2007). A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers. Computer Communications, 30: 2201-2212.

Tsai, C.F. and Lin, C.Y. (2010). A triangle area-based nearest neighbors approach to intrusion detection. Pattern Recognition, 43(1):222-229.

Tsang, C.H., Kwong, S., and Wang, H. (2007). Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection. Pattern Recognition, 40:2373–2391.

Upadhyaya, S., Chinchani, R., and Kwiat, K. (2001). An analytical framework for reasoning about intrusions. In Proceedings of the IEEE Symposium on Reliable Distributed Systems, New Orleans, LA pp. 99–108.

Vapnik, V. (1995). The Nature of Statistical Learning Theory,SVM. Springer-Verleg.

W. Lee, S.J. Stolfo, P.K. Chan, E. Eskin, W. Fan, M. Miller, S. Hershkop and J. Zhang. (2001). Real time data mining-based intrusion detection, in: Proceedings of the DARPA Information Survivability Conference & Exposition II, Anaheim, USA, pp. 89–100.

Wu, T.F., Lin, C.J., and Weng, R.C. (2004). Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research, 5: 975-1005.

Xiang, C., Chong, M.Y., and Zhu, H.L. (2004). Design of multiple-level tree classifiers for intrusion detection system. In Proceedings of 2004 IEEE Conference on Cybernetics and Intelligent Systems, Singapore, pp. 872–877.

Xiang, C., Yong, P.C., and Meng, L.S. (2008). Design of multiple level hybrid classifier for intrusion detection system using Bayesian clustering and decision tree. Pattern Recognition Letters, 29: 918-924.

Xindong, W., Vipin, K., and Ross, J.Q. (2008). Top 10 algorithms in data mining. Knowledge Information Systm. 14:1–37.

Yang Li and Li Guo. (2007). An active learning based on TCM-KNN algorithm for supervised network intrusion. Computer and Securtiy. 26: 459-467.

Zalik, K.R. (2008). An efficient *K*-Means clustering algorithm. Pattern Recognition Letter 29:1385-1391.

Zhang, C., Jiang, J., and Kamel, M. (2005). Intrusion detection using hierarchical neural network. Pattern Recognition Letters, 26: 779–791.

Zhang, H. and Su, J. (2008). Naive Bayes for optimal ranking. Journal of Experimental and Theoretical Artificial Intelligence, 20:79-93.