An Analysis of Decision Tree Based Intrusion Detection System

Yi Yi Aung, Myat Myat Min
University of Computer Studies, Mandalay
yiyiaung123@gmail.com, myatiimin@gmail.com

Abstract

Intrusion detection is the process called indentifying intrusions. The action of entering to a system without permission is called intrusion. With the improving advanced technology of mobile devices such as smart phones, tablet, smart devices, other computing devices, the number of network users are increasing more and more. Hence, security on network is very important for all net consumers. IDS are fundamental part of security boundary. So, they are now considered as a mandatory safety mechanism for critical networks. There are many traditional techniques of intrusion detection. In the research of traditional intrusion detection technology analysis, the statistical model for the establishment of the regulatory basis, management and aggression capability and so on there are still some disadvantages and disabilities, because actual test results cannot meet the requirements. Current methods used in IDS are many. Each method has advantage and disadvantage. Intrusion detection can also be seen as a classification problem. In this research we use K-means and C4.5 algorithms. This paper presents the comparison of intrusion detection by using hybrid data mining methods and a single method. The purpose of this paper is to show the differences of time complexity between hybrid data mining method and a single method. This model is verified using KDD'99 data set. Experimental result clearly shows hybrid methods can reduce model training time while maintaining the higher detection rates than using single method.

Keywords- Intrusion Detection System, KDD'99 dataset, K-means, C4.5

1. Introduction

People want to keep their possessions secure. So they consider many ways to secure their possessions and then invented many software and hardware devices to protect their belongings. There is no secure system in world but we must consider and protect our system as much as we can.

The computer system controls a large amount of data over the network, so data communications should be secure enough for data transceivers. In the previous day's firewall, data encryption, antivirus is used to prevent unauthorized access to the network system. There is a technique known as the intrusion detection system that will be used to monitor unwanted user data over a network. [4]

Cyber security is a critical topic for both researchers and practitioners as successful cyber attacks can result in severe costs due to losses of confidentiality, integrity or availability. Various security mechanisms have been suggested for detecting cyber attacks such as intrusion detection system. Intrusion detection system is sometimes called classification problems.

Intrusions in a computer system are the activities that infringe the system security policy. The process of monitoring and analyzing the activities occurring in the process of Intrusion Detection process is in a deep way. [1]

Detection of intrusions identifies an unauthorized individual to use a computer, and identifies an authorized individual who abuse their power. The intrusion Detection System is an important defense tool for network security. By analyzing audit data, the Intrusion Detection System tells the administrator to take appropriate action to protect the application system from further attacks when the system is in unsafe condition. Because network intrusion is a set of human behavior and user behavior, it can be divided into normal and abnormal behavior. [2]

Many security experts and researchers have proposed and implemented different strategies to defend the computer system from attacks. Among them the intrusion detection systems (IDSs) can be specifically designed to identify attacks that can target computers or networks and resources. IDS have two main components: data audit component, sensors or log files, which monitors / collects data on system behavior; a component detection method that analyzes data that is observed/collected to detect malicious activity. In terms of audit components, IDS is classified as host-based (HIDS) or network-based (NIDS). Host-based IDS detects attacks on computer system by monitoring mainly operating system activity. Network-based IDS detects nodes connected to the network by monitoring TCP/IP events. In terms of detection methods, IDS is further classified as signature or anomaly tracking system. Signature-based systems monitor traffic to known attack patterns (signatures), similar to virus scanners that protect personal computers. Signature-based IDSs efficiently detect existing threats but always miss new threads.

Finally, KDD CUP 1999 dataset is used to verify the effectiveness of our method. The experimental results

ISBN 978-99971-0-381-9 © 2017 ICAIT

show that the collaborative intrusion detection and customization methods proposed in this paper are superior to C4.5 detection in tracking accuracy and tracking efficiency.

2. Literature Survey

Many data mining algorithms are applied to intrusion detection. This is divided into general offline algorithms and inline incremental algorithms. Most researchers have focused on off-line intrusion detection using a well-known KDD99 dataset and verified the development of IDS. The KDD99 dataset is a statistically preprocessed dataset that has been available since 1999. [17]

Classification is one of the important functiona of data mining. It performs its task by classifies the data into different categories using the algorithm type. This classification has wide range of applications in the field of network intrusion detection. It categories network patterns as normal or attack to identify malicious activities occurring in the network. In this paper, researcher is analyzing classification algorithms using NSL KDD 99 dataset. [9]

Intrusion detection is one of the difficult problems encountered by the modern network security industry. Data mining can play a important role in system development. Data mining is a technique for extracting important information from a huge data repository. In order to detect intrusion, the traffic created in the network can be broadly categorized into following two categories-normal and anomalous. In this proposed paper, several classification techniques and machine learning algorithms have been considered to categorize the network traffic. The comparison of data mining algorithms has been performed using WEKA tool and listed below according to certain performance metrics. Simulation of these classification models has been performed using 10-fold cross validation. For this simulation, they used NSL-KDD based data set in WEKA. [3]

With the rapid development of computer networks during the past decade, security has become a key issue for computer systems. It is the IDS which protect to our computer network. Different classification and clustering algorithms have been proposed in recent year for the implementation of intrusion detection systems. In this paper, multiple algorithms are analyzed to find the optimal algorithm. At last the optimal algorithms Random Forest and DB Scan are occurred. [5]

In this paper, the various intrusion-detection-system techniques and their application on the basis of decision tress also discussed that are available and on which various researches have made. Some detection approaches that are applied for the intrusion detection are focused on some specific methods. In this paper various intrusion detection approaches are analyzed for detection of intrusion by the use of decision tree algorithm. [6]

Intrusion is an act that violates the security policy of the system. The purpose of this research paper is to explain the method / technique used for intrusion detection based on the concept of data mining and the structure designed for it. This survey document outlines the intrusion detection process and the data mining methods and methods to facilitate the frames developed using these concepts. [10]

Since the ready-made data mining algorithms is offered, intrusion detection based on the data mining has improved rapidly. It advances in the ability to hold enormous data, but it also has troubles like, for instance, searching for more helpful data mining algorithms , how to progress the correct rate of intrusion detection, and etc. These can be the topic for future study; meanwhile they also need lots of effort and experiments to develop a system that is more effective and more suitable. There are many types of approaches in intrusion detection, in which that based on the data mining becomes the hot topic in the current intrusion detection methodology. However, data mining is still in its developing stage, so more thorough study needs to be done. A brief survey of the IDS in the data mining field is given in this paper. [12]

In this paper, they present an intrusion detection system using J-48 and Naïve Bayes for classification. To implement and classify of the system they used KDD 99 dataset and their University’s traffic. The principal challenge in intrusion detection is to obtain high detection rate. From this paper’s experimental result shown as single classifier is not sufficient to obtain the high result and feature selection is the most important to detection ratio also showed that the effectiveness of J-48 is comparable to the Naïve Bayes. [13]

This paper draws the conclusions on the foundation of implementations performed using various data mining algorithms. Combining more than one data mining algorithms may be used to eliminate disadvantages of one another. Thus a combining approach has to be made while selecting a mode to apply intrusion detection system. Combining a number of qualified classifiers lead to a better performance than any single classifier. [15]

3. Intrusion Detection System and Data Mining

Monitoring user activity on the network and categorizing malicious and normal activity is called intrusion detection. The system used for this purpose is called Intrusion Detection System (IDS). Intrusion detection systems are combinations that perform software or hardware, or automated processes that track and analyze events. In general, IDS monitors and records computer system events, performs to determine if an event is a security incident, alerts potential threat to security employees, and generates event reports. Every time an intruder attempts to compromise the confidentiality, integrity or availability of the network or system, IDS monitors and detect illegal activities, prohibit legitimate users to access resources or computer to system services. It also takes appropriate predefined expectations into account and performs appropriate actions.

The intrusion detection techniques can be defined as a system that identifies and deals with malicious use of computer and network resources. The IDS using its detection techniques tracks the user available on the network and traces the activities being carried out. The audit data after being traced are compared to the known awful records and an alarm is set whether the similarities of the two are above some predefined threshold. The user is then accordingly distinguished as normal or illicit user.

The IDS techniques on the basis of their detection process can be categorized into two methodologies:

Misuse Detection technique:

The misuse detection technique uncovers the intruder activities based on the extensive knowledge of known patterns provided by human experts. The detection process involves matching features through the attacking feature library and confirming the attack incidents. The key advantage of misuse detection system is that once the patterns of known intrusions are stored, future instances of these intrusions can be detected effectively and efficiently. The detection process though can even catch the negligible intrusive activities and generates the much fewer false alarms but still is unable to detect novel or unknown attack. [11]

Anomaly Detection technique:

Anomaly detection technique analyses the intrusive activity and identifies the new intrusion types according to the deviation of a computer from its normal usage. If the divergence is much enough then the user activity is considered as abnormal. The key advantages of anomaly detection systems are that they can detect unknown intrusion since they require no a priori knowledge about specific intrusions. Anomaly detection although reveals the new trends of intrusions still lacks the detection of negligible intruder activities and also generates higher false alarm rate.

After focusing on the IDS techniques, the IDS based on its analysis and audit data storage unit are of two types:

Host-based IDS (HIDS):

HIDS is a host based detection approach in which a system collects the data as the records of various activities of host including event logs, system logs etc. As the system monitors only the host or agent it determines the awfulness more accurately. As everything is on the host there is no need of installing additional hardware or software but still redundancy is one of the important issue especially when we desire to install the system for a network and we should have a HIDS for each host. Since individual monitoring system for each host is needed, the efficiency in terms of speed decreases and the system cost increases.

Network-based IDS (NIDS):

NIDS is a network based approach in which the system in place of collecting data from a particular host/agent directly collects it from the network monitored in form of packets. It provides better security against DoS attacks as compare to HIDS. Mostly NIDS are operating system independent and are easy to deploy. The system does not need to be installed on multiple monitoring systems hence are less expensive but lacks accuracy due to loosing of some data during the detection process. Also for large scale network scalability is still a problem. [7] [14]

Data mining (DM), also called Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns using association rules. It is a fairly recent topic in computer science but utilizes many older computational techniques from statistics, information retrieval, machine learning and pattern recognition. Data mining techniques can be differentiated by their different model functions and representation, preference criterion, and algorithms. The main function of the model that we are interested in is classification, as normal, or malicious, or as a particular type of attack.

The use of data mining techniques in IDSs usually implies analysis of the collected data in an offline environment. There are important advantages in performing intrusion detection in an offline environment, in addition to the real-time detection tasks typically employed.

There are several reasons why data mining approaches plays a role in these three domains. First of all, for the classification of security incidents, a vast amount of data has to be analyzed containing historical data. It is difficult for human beings to find a pattern in such an enormous amount of data. Data mining, however, seems well-suited to overcome this problem and can therefore be used to discover those patterns. [8]

4. Methodology

This paper involves discussion of the two algorithms of data mining classification approaches, K-means and C4.5.

K-means algorithm:

The K-means algorithm is one of the most popular methods of clustering analysis that aims to partition ‘n’ data objects into ‘k’ clusters in which each data object belongs to the cluster with the nearest mean. It uses Euclidean metric as a similarity measure. The important properties of k-means algorithms are efficient in processing large datasets and it can work only on numerical values.

The basic algorithm of k-means is:

1. Select k objects as initial centroids
2. Assign each object to the closest centroids.

3. Recalculate the centroid of each cluster.
4. Repeat steps 2 and 3 until centroids do not change.

C4.5 algorithm:

Decision trees can be used as misuse intrusion detection as they can learn a model based on the training data and can predict the future data as one of the attack types or normal based on the learned model. It works well with large data sets. It constructs easily interpretable models, which is useful for a security officer to inspect and edit. It can also be used in the rule-base models with minimum processing.

The system flow diagram of this paper by using the K-means and C4.5 can be seen in figure 1.

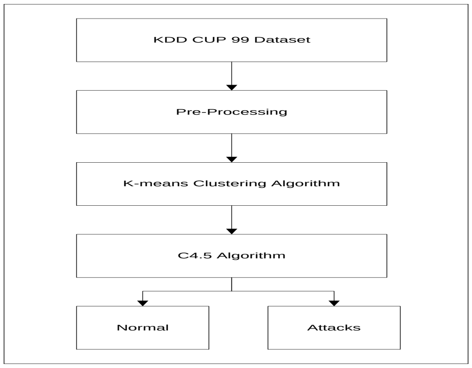

Figure 1. System Flow Diagram

5. Experiment and Result Analysis

To facilitate the experiments, we used Eclipse Java and Weka tool to implement the algorithms on a PC with 64-bit window 7 operating system, 4GB RAM and a CPU of Intel core i3-4010U CPU with 1.70GHz.

5.1. Data Selection

The experimental analysis is done by considering the typical dataset for intrusion detection named as KDD CUP'99. It is the most widely preferred dataset especially formulated for examining the newly implemented intrusion detection models. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment. The data set includes 42 attributes classifying the data records into normal or a type of attack. The Table 1 notify about the 41 types of attributes in the KDD CUP'99 dataset categorized into 5 major attack classes under which they fall. [16]

Table 1: Various attacks and their respective categories

Class	Known Attacks Subclass
DoS	back, land, Neptune, pod, smurf, teardrop
Probe	ipsweep, nmap, portsweep, satan
U2R	buffer_overflow, loadmodule, perl, rootkit
R2L	ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster

Kddcup'99 dataset have two variations of training dataset; one is a full training set having 5 million connections and the other is 10% of this training set having 494021 connections. Since the whole dataset is vast, the experiment has been performed on its smaller version that is 10% of KDD.

5.2. Result and Analysis

The analysis is performed by using K-means and C4.5 algorithms. We use K-means algorithm to generate heterogeneous dataset to nearly homogeneous dataset. Then we apply C4.5 algorithm to know the intrusions and normal traffic. For the experiment, we have run the simulation with the five different sizes of partition by using k-means algorithm and then use C4.5 algorithm for detection. And also we have run the dataset without using k-means algorithm. Then, we can compare the accuracy of two approaches. One approach is using k-means and C4.5 algorithm and the other approach is using only C4.5 algorithm. The comparison of these two approaches can be seen from table 2 to table 5.

The training time of C4.5 algorithm based on K-means is 3546.66 seconds, while that of only C4.5 algorithm is 7969.19 seconds in 10 fold cross validation. The training time of C4.5 algorithm based on K-means is 3198.51 seconds, while that of only C4.5 algorithm is 7181.49 seconds in 66-34 percentage validation. This indicates that the collaborative intrusion detection based on K-means and C4.5 is better than a single C4.5 algorithm in the training time.

Table 2. Comparison Testing Results of Two Approaches for 10 fold

dataset	k-means	C4.5	Correct instances	Correct percent	Incorrect instances	Incorrect percent
10% P1	Y	Y	108826	99.9825	19	0.0175
10% P2	Y	Y	23495	99.8597	33	0.1403
10% P3	Y	Y	280798	100	0	0
10% P4	Y	Y	78656	99.8718	101	0.1282
10% P5	Y	Y	2066	98.71	27	1.29
Total			493841		180	
10%kdd	N	Y	493823	99.9599	198	0.0401

Table 3. Comparison Testing results of Two Approaches for 10 fold with time complexity

dataset	k-means	C4.5	Total instances	Time to build model (sec)
10% P1	Y	Y	108845	907.86
10% P2	Y	Y	23528	37.79
10% P3	Y	Y	280798	2036.96
10% P4	Y	Y	78757	563.91
10% P5	Y	Y	2093	0.14
Total				3546.66
10%kdd	N	Y	494021	7969.19

Table 4. Comparison Testing Results of Two Approaches for 66-34 percentage

dataset	k-means	C4.5	Correct instances	Correct percent	Incorrect instances	Incorrect percent
10% P1	Y	Y	37000	99.9811	7	0.0189
10% P2	Y	Y	7984	99.8	16	0.2
10% P3	Y	Y	95471	100	0	0
10% P4	Y	Y	26736	99.8469	41	0.1531
10% P5	Y	Y	702	98.5955	10	1.4045
Total			167893		74	
10%kdd	N	Y	167876	99.9458	91	0.0542

Table 5. Comparison Testing results of Two Approaches for 66-34 with time complexity

dataset	k-means	C4.5	Total instances	Time to build model (sec)
10% P1	Y	Y	37007	864.42
10% P2	Y	Y	8000	35.89
10% P3	Y	Y	95471	1722.84
10% P4	Y	Y	26777	575.23
10% P5	Y	Y	712	0.13
Total				3198.51
10%kdd	N	Y	167967	7181.49

The total correctly classified instances based on K-means and C4.5 are 493841 while that of instances based on only C4.5 is 493823 in 10 fold cross validation. The total correctly classified instances based on K-means and C4.5 are 167893 while that of instances based on only C4.5 is 167876 in 66-34 percentage validation. This shows that hybrid method can correctly classify than a single method.

6. Conclusion

Experimental results show that the optimized and adaptive collaboration intrusion detection model based on K-means and C4.5 is superior to the detection system with a single C.5 algorithm in the detection accuracy and time efficiency.

Further work will be directed to experimental research of the data mining methods, approaches and algorithms by using real network data.

7. References

[1] K.A. Al-Enezi, I.F. Al-shaikhli, A.R. Al-kandari, L. Z. All-Tayyar, "A Survey of Intrusion Deteection System using Case Study Kuwait Governments Entiteis ", 3rd International Conference on Advanced Computer Science Applications and Technologies, 2014.

[2] L. Teng, S. Teng, F. Tang, H. Zhu, W. Zhang, D. Lin and L. Liang, "A Collaborative and Adaptive Intrusion Detection Based on SVMs and Decision Trees", IEEE International Conference on Data Mining Workshop, 2014.

[3] S. Choudhury and A. Bhowal, "Comparative Analysis of Machine Learning Algorithms along with Classifiers for Network Intrusion Detection", International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2015,pp.89-95.

[4] S.H. Vasudeo, P.P. Patil and R.V. Kumar, “IMMIX-Intrusion Detection and Prevention System”, International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2015,pp.96-101

[5] P.S. Rath, M. Hohanty, S. Acharya and M. Aich, “Optimization of IDS Algorithms Using Data Mining Technique”, Proceeding of 53rd IRF International Conference, Pune, India, 2016,ISBN:978-93-86083-01-2.

[6] S. Singh and S. Jain, “A Comparative Analysis of Decision Tree Based Intrusion Detection System”, International Journal of Modern Trends in Engineering and Research, Scientific Journal Impact Factor (SJIF), 2016, ISSN (Online):2349-9745.

[7] M. Dhakar, N. Chaurasia and A. Tiwari, “Analysis of K2 based Intrusion Detection System”, Current Research in Engineering, Science and Technology (CREST) Journals, 2013, ISSN 2320-706X.

[8] R. Patel, A. Thakkar and A. Ganatra, “A Survey and Comparative Analysis of Data Mining Techniques for Network Intrusion Detection Systems”, International Journal of Soft Computing and Engineering (IJSCE), March 2012, ISSN: 2231-2307, Volume-2, Issue-1.

[9] M.P. Bhoria and Dr.K. Garg, “An Imperial learning of Data Mining Classification Algorithms in Intrusion Detection Dataset”, International Journal of Scientific & Engineering Research, June-2013, Volume 4, Issue 6, ISSN 2229-5518.

[10] R. Venkatesan, R. Ganesan and A.A.L. Selvakumar, “A Comprehensive Study in Data Mining Frameworks for Intrusion Detection”, International Journal of Advanced Computer Research, December-2012, Volume-2 Number-4 Issue-7, ISSN (print): 2249-7277 ISSN (online): 2277-7970.

[11] J. Cannady and J. Harrell, “A Comparative Analysis of Current Intrusion Detection Technologies”.

[12] L.S. Parihar and A. Tiwari, “Survey on Intrusion Detection Usingn Data Mining Methods”,IJSART, January-2016, Volume-2 Issue-1 ISSN (online): 2395-1052.

[13] Ugtakhbayar.N , Usukhbayar.B and Nyamjav.J, “An approach to detect TCP/IP based attack”, IJCSNS International Journal of Computer Science and Network Security, April-2016, Vol-16 No-4.

[14] E. Bloedorn, A.D. Christiansen, W. Hill, C. Skorupka, L.M. Talbot and J. Tivel, “ Data Mining for Network Intrusion Detection: How to Get Started”.

[15] TR. Patel, A. Thakkar and A. Ganatra, “A Survey and Comparative Analysis of Data Mining Techniques for Network Intrusion Detection Systems”, International Journal of Soft Computing and Engineering (IUSCE), March-2012, Vol-2, Issue-1, ISSN: 2231-2307.

[16] M. Dhakar and A. Tiwari, “A New Model for Intrusion Detection based on Reduced Error Pruning Technique”, I.J. Computer Network and Information Security, 2013,22,51-57.

[17] A.A. Nasr, M.M. Ezz and M.Z. Abdulmageed, “Use of Decision Trees and Attributional Rules in Incremental Learning of an Intrusion Detection Moedl”, International Journal of Computer Networks and Communications Security, July-2014, Vol-2,No-7, 216-224, ISSN 2308-9830.