An Analysis of Rule and Decision Tree Based Intrusion Detection System

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

Abstract

We are living in 21st century wherein the number of internet of things is competing with increasing population. Security is becoming a major concern for information technology over network. Therefore, people use technology to overcome every problem that comes in the network intrusions. Now many researchers and developers are trying to protect networks from various attacks but at the same time raise many questions, confusions and conflicts regarding their protecting technology. Because each approaches have advantages and disadvantages in detection. The system use rule based data mining techniques in intrusion detection system for network. And also it compares the approaches of rule based and tree based intrusion detection system by using 10 % kddcup'99 dataset. For rule based approach, we use K-means and JRip algorithms to classify internal and external security threats and attacks. For decision tree based approach, we use K-means and C4.5 algorithms to compare between detection methods.

1. Introduction

During the past few years, security of computer network has become main stream in most of everyone's lives. An intrusion detection system (IDS) is a type of security software designed to automatically alert administrators when someone or something is trying to compromise information system through malicious activities or through security policy violations. An IDS works by monitoring system activity through examining vulnerabilities in the system, the integrity of files and conducting an analysis of patterns based on already known attacks. It also automatically monitors the internet to search for any of the latest threats that could result in a future attack. The IDS is also a listen-only device. The IDS monitors traffic and report it result to an administrator. For controlling intrusions, intrusion detection systems are introduced [2, 14, and 18].

Due to the effective data analysis method, data mining is introduced in intrusion detection system.

2. Literature Review

Jaina Patel and Mr. Krunal Panchal used anomaly based intrusion detection technique with signature based intrusion detection technique by dividing into two stages. Firstly, the signature base IDS SNORT is used to generate alerts for anomaly data. And then, data mining techniques "K-means +CART" is used to cascade k-means clustering and CART (Classification and Regression Trees) for classifying normal and abnormal activities. They aimed to maximize the effectiveness in identifying attacks and achieve high accuracy rate as well as low false alarm rate [6].

Prof. S. V. Peterraj and S. P. P. Mary attempted to evaluate, categorize, compares and summarizes the performance of data mining techniques to detect the intrusion and summarized the results of various researches on suitability of data mining techniques for intrusion detection [7].

A. Narayan M. and T. J. Parvat proposed a model using a multi-layer Hybrid Classifier to estimate whether the action is an attack or normal data. Firstly, a misuse detection model is built based on the C4.5 decision tree algorithm and then the normal training data is decomposed into smaller subsets using the model. Next, multiple one-class Naïve Byes algorithm models are created for the decomposed subsets. Hybrid Classifier is used as a preprocessor of Intrusion Detection System to reduce the feature and training time [8].

Neethu B described PCA for feature selection with Naïve Bayes for classification in order to build a network intrusion detection system. The experimental results showed that the proposed approach was very accurate with low false positive rate and takes less time in comparison to other existing approaches while building an efficient network intrusion detection system [9].

Saranya. V and Amsaveni. R presented three classification techniques in that they selected two

classifiers from each technique. Based on the attack classifying for DoS attack JRip performed well and for Probe attack Naïve bayes performed well and U2R attack bayesNet performedd well and for oneR performed well. By considering overall performance JRip rule based classifier performs well [12].

M. R. Kabir, A. R. Onik and T. Samad presented a wrapper approach for intrusion detection. In this framework feature selection technique eliminate the irrelevant features to reduce the time complexity and build a better model to predict the result with a greater accuracy and Bayesian network works as a base classifier to predict the type of attack. Their model performed better than other leading state-of-the-arts models such as KNN, Boosted DT, Hidden NB and Markov chain. The NSL-KDD is used as benchmark data set with weka library functions in the experimental setup [13].

Upendra and Y. K. Jain reported on the empirical evaluation of five machine leaning algorithms such as J48, BayesNet, OneR, NB and ZeroR using ten performance criteria such as accuracy, precision, recall, incorrectly classified instances, etc. J48 with an accuracy rate of approximately 99% was found to perform much better at detecting intrusions than others based on the experiments done in the paper and their corresponding results. Therefore, they stated that J48 classifiers shows better performance for all classes (Normal, DOS, R2L, U2R, Prob) [16].

Neethu B discussed about the combinational use of two machine learning algorithms called Principal Component Analysis and Naïve Bayes classifier. The dimensionality of the dataset was reduced by using the principal component analysis and the classification of the dataset into normal and attack classes was done by using Naïve Bayes Classifier. The comparison of the results with or without dimensionality reduction was also done [17].

K. Umamaheswari and S. Janakiraman evaluated performance of a comprehensive set of classifier algorithms using KDD99 dataset. They then used WEKA to bring out an extensive performance comparison among the most popular classifier algorithms such as Random Forest, Random Tree, J48, etc. In their paper, they conclude Random Tree algorithm produces better accuracy compared with other algorithms [18].

Intrusion Detection System (IDS) was said to be more effective when it had both high intrusion detection rate and low false alarm. But current IDS when implemented using data mining approach like clustering, classification alone are unable to give 100% detection rate hence lack effectiveness. In order to overcome these difficulties of the existing systems, many researchers implemented intrusion detection systems by integrating clustering and classification approach like K-means and Fuzzy logic, K-means and genetic algorithm, some of the researcher also tried use of Decision tree and Neural Network to detect unknown attacks. Therefore, P. R. Patil, Y.Sharma and M. Kshirasagar analyzed Hybrid system analysis's which were implemented by using the benchmark dataset compiled for the 1999 KDD intrusion detection contest, by MIT Lincoln Labs [19].

Network Intrusion Detection System is a latest kind of defense technology which is one of the vibrant areas in network security. A variety of intrusion detection approaches be present to resolve this severe issue but the main problem is performance. S. K. Jonnalagadda and R. P. Reddy I analyzed a detailed survey of important techniques based on intrusion detection. Also the classification of the techniques based on neural network, K-means, hybrid techniques, support vector machine etc., was provided. For comprehensive analysis, detection rate, time and false alarm rate from various research papers had been taken [20].

Y. Y. Aung and M. M. Min used K-means and C4.5 algorithms to detect intrusions or not. The purpose of their paper was to provide an intrusion detection system by using hybrid data mining methods. This model was verified using KDD'99 dataset. Experimental results clearly showed that the model provided higher detection rates and lower training time [15].

3. Intrusion Detection System

Intrusion Detection System (IDS) is the level of security used to detect continuous interference in information systems. Traditionally, intrusion detection is based on extensive knowledge of security professionals, especially those familiar with secure computer systems. To facilitate this dependency, many data mining and machine learning methods are introduced to detect intrusion. Intrusion detection is one of the most difficult issues in the latest cyber security industry. The anomaly of the anomalies on the internet (often referred to as intrusion) causes security issues that lead to the successful implementation of an anomaly detection system known as intrusion detection system (IDS). Intrusion

can be defined as phishing attempts to access a protected system or network. Intrusion detection is an action process for detecting suspicious activity on a network or device. Intrusion Detection System (IDS) is an important discovery used to protect the integrity of data and system availability from attacks [1].

3.1. Intrusion Detection Process

Intrusion detection on the basis of their detection process are categorize into Misuse / Signature-based intrusion detection and Anomaly-based intrusion detection.

3.1.1. Misuse Detection

Misuse detection compares the user activities to the known intruder activities on web. The idea of abusive detection is to show off attacks as templates and signatures, and to prevent and block such attacks in the future. IDS searches for specific signatures and if matching is found, the system will generate a warning signal indicating an intrusion. Although misuse based intrusion detection system can easily detect know attacks by using predefined signatures, it is impossible to detect new attacks, for which no pattern is available [17].

3.1.2. Anomaly Detection

Anomaly intrusion detection identifies deviations from the normal usage behavior patterns to identify the intrusion. This technology is based on the evidence of traffic anomaly. It evaluates the deviation between user activity and normal behavior and assumes that activity is intrusive if the deviation exceeds the specified threshold. This threshold concept abnormality can detect new infiltrations in addition to previously known invasions. However, anomalies can detect new invasions, but coercion to engage in limiting factors increases the false positive rate [1, 17].

3.2. Intrusion Detection Approaches

On the basis of the data analyzed and stored it is classified into Host-based Intrusion Detection System (HIDS) and Network-based Intrusion Detection System (NIDS).

3.2.1. Host-based Intrusion Detection System

Host-based IDS analyze server-related audit resources, such as operating system audit logs, system logs, application logs. Software installed in the system to protect the system from intruders. Audit data analyzed will be collected from the server (or system) on the network. Because the HIDS operating system depends on the operating system, it should be planned before implementation and you can effectively detect buffer overflow attacks [1, 2].

3.2.2. Network-based Intrusion Detection System

Network-based IDS analyze network packets obtained on the network. Intrusion detection was detected in the NIDS discovery program on the network. NIDS collects data as packets directly from the network and is analyzed for intrusion detection. Protocols and content cannot be scanned if traffic is encrypted independently of the operating system. This helps to better protect against denial of service attacks [1, 2].

4. Data Mining for Intrusion Detection System

Data Mining is identified as a solution to handling the analysis of data due to its adaptability and validity and it is now used extensively used for network security purposes. As an application of machine learning, data mining holds a very significant position in intrusion detection, presenting methods of predicting future patterns based on past experiences. The practice of data mining has been applied to many fields, such as sales, healthcare, medical, finance, multimedia and most importantly intrusion detection. There are many types of algorithms that may be used to mine audit data. Data mining can play a massive role in the development of a system which can detect network intrusion [3].

Clustering is a process of splitting data into clusters based on the features of the data. This clustering partitions data into groups of similar objects. Each member within the cluster is similar to one another. The classification algorithm may be able to predict new unseen data classifying it by using pre-existing information [3]. Standard classification algorithms do not perform well when the computer intrusions are much rarer than normal behavior. In such scenarios, researchers have developed special algorithms and applied to intrusion detection problems. Moreover, the classification accuracy of the existing algorithms or techniques has to be improved as it is very difficult to detect new attacks. Classifier is a challenge to build and efficient

intrusion detection system [10]. In the system, we use hybrid data mining technique to detect normal and attacks.

4.1. K-means Algorithm

K-means clustering algorithm takes set of data points and k clusters as input and places the k-centroids in random location in space. Then find the nearest centroid for each record in the dataset. This is done by making use of Euclidian distance formula to find the distance between data points and every cluster centroid. Then choose the cluster which as minimum distance to the centroid. Recalculate the centroids position by considering all the data points that belongs to cluster. The process is repeated until the centroids position is unchanged [2, 20].

4.2. JRip Algorithm

JRip (RIPPER) is one of the most popular algorithms; it has classes that are examined in increasing size. It also includes set of rules for class is generated using reduced error Jrip (RIPPER). Proceed by treating examples of judgments made in training data as a class, and finding rules that covers all the members of the class. Then it proceeds to the next class and repeats the same action, repetition is done until all classes have been covered [4, 5, and 11].

4.3. System Architecture

The system flow diagram of this paper by using the K-means and JRip can be seen in Figure 1.To implement the Intrusion Detection System we apply our detection model on the 10 percent of KDDCUP'99 dataset.

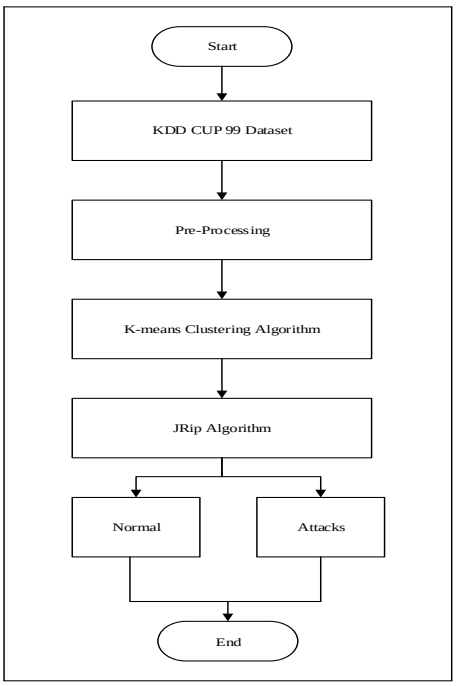

Figure 1. System Flow Diagram

5. Experimental Results and Discussion

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. Data come from MIT Lincoln laboratory of KDDCup99 data set. We select 10% data set which contains 494021 Connection records, each record has total of 41 characteristics, 7 symbolic field and 34 numeric fields to experiment because the data sets are very huge. This data set contains four types of intrusions: DoS, Probe, U2R and R2L and also contain normal samples as shown in table 1[3]. Research activities in IDS are still using the KDD Cup 99 dataset for analyzing and exploring new approaches for better IDS.

Table 1. Class-wise Attack on KDD Dataset

Class of attack	Attack Name
Normal	Normal
DoS	Neptune, Smurf, Pod, Teardrop, Land, Back
Probe	Ipsweep, nmap, satan, portsweep
R2L	ftp_write, guess_passwd,imap, multihop, phf, spy
U2R	Perl, buffer_overflow, rootkit, loadmodule

Kddcup'99 dataset have two variations of training dataset; one is full training set having 5 million connections and the other is 10% of this training set having 494021 connections. These dataset can be seen in Table 2. We use 10% data set of KDD CUP 99 to correctly classify the normal and intrusions in the data set. The experimental results of the system are shown from Table 3 to Table 6. The testing results of K-means and JRip algorithms based on 10 fold cross validation can be seen in table 3 and table 4. And also the results of based on 66-34 percent validation can be seen in table 5 and table 6.

Table 2. Number of Instances in KDD and 10%KDD

Class	Whole KDD	10% KDD
DoS	3883370	391458
Probe	41102	4107
U2R	52	52
R2L	1126	1126
Normal	972780	97278
Total	4898430	494021

Table 3. Testing Results for 10 Fold Cross Validation

Dataset	K-means	JRip	Correctly Classified Instances	Correct Instances Percentages	Incorrectly Classified Instances	Incorrect Instances Percentages
10% P1	Y	Y	108830	99.9862	15	0.0138
10% P2	Y	Y	23500	99.881	28	0.119
10% P3	Y	Y	280798	100	0	0
10% P4	Y	Y	78683	99.906	74	0.094
10% P5	Y	Y	2063	98.5667	30	1.4333
Total	Y	Y	493874		147	

Table 4. Testing Results for 10 Fold Cross Validation with Time Complexity

Dataset	K-means	JRip	Total Instances	Time to Build Model (Sec)
10% P1	Y	Y	108845	1809.53
10% P2	Y	Y	23528	97.94
10% P3	Y	Y	280798	1293.5
10% P4	Y	Y	78757	2576.09
10% P5	Y	Y	2093	1.66
Total	Y	Y	494021	5778.72

Table 5. Testing Results for 66-34 Percent Validation

Dataset	K-means	JRip	Correctly Classified Instances	Correct Instances Percentages	Incorrectly Classified Instances	Incorrect Instances Percentages
10% P1	Y	Y	37003	99.9892	4	0.0108
10% P2	Y	Y	7987	99.8375	13	0.1625
10% P3	Y	Y	95471	100	0	0
10% P4	Y	Y	26741	99.8656	36	0.1344
10% P5	Y	Y	705	99.0169	7	0.9831
Total	Y	Y	167907		60	

Table 6. Testing Results for 66-34 Percent Validation with Time Complexity

Dataset	K-means	JRip	Total Instances	Time to Build Model (Sec)
10% P1	Y	Y	37007	1760.96
10% P2	Y	Y	8000	83.83
10% P3	Y	Y	95471	1227.63
10% P4	Y	Y	26777	2459.85
10% P5	Y	Y	712	1.33
Total	Y	Y	167967	5533.6

This paper also describes the comparison results of decision tree based intrusion detection system and rule based intrusion detection system in table 7 and table 8.

Table 7. Comparison Results for 10 Fold Cross Validation

Data set	Method	Total Instances	Correct Instances	Incorrect Instances	Time Taken To build Model (sec)
10% kdd	K-means + C4.5	494021	493841	180	3546.66
10% kdd	K-means + JRip	494021	493874	147	5778.72

Table 8. Comparison Results for 66-34 Percent Validation

Data set	Method	Total Instances	Correct Instances	Incorrect Instances	Time Taken To build Model (sec)
10% kdd	K-means + C4.5	167967	167893	74	3198.51
10% kdd	K-means + JRip	167967	167907	60	5533.6

In analysis of 10 fold cross validation, the correctly classified instance records of decision tree based approach is 493841 records while that of rule based approach is 493874 records. And the incorrectly classified instance records of decision tree based approach is 180 records while that of rule based approach is 147 records. The time needed to train the model for decision tree based approach is 3546.66 seconds while that of rule based approach is 5778.72 seconds.

And also, in 66-34 percent validation, the correctly classified instance records of decision tree based approach is 167893 records while that of rule based approach is 167907 records. And the incorrectly classified instance records of decision tree based approach is 74 records while that of rule based approach is 60 records. The time needed to train the model for decision tree based approach is 3198.51 seconds while that of rule based approach is 5533.6 seconds.

6. Conclusion

This paper proposes a hybrid intrusion detection framework. This framework use two data mining techniques (i.e. K-means and JRip) to detect normal and attacks. And also this paper describes the comparison results of decision tree based IDS (i.e. K-means and C4.5) and rule based IDS. Experimental results show that the accuracy of JRip algorithm based on K-means is good than that of C4.5 algorithm based on K-means. But the model training time of JRip algorithm based on K-means algorithm is more take time than that of C4.5 algorithm based on K-means algorithm to train the model.

References

[1] M. Dhakar and A. Tiwari, "A Novel Data Mining based Hybrid Intrusion Detection Framework", Journal of Information and Computing Science, Vol. 9, No. 1, pp. 037-048, ISSN 1746-7659, England, UK, 2014.

[2] Vibha Rao, "A Clustering Algorithm for Intrusion Detection using Hybrid Data Mining Technique", International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, National Conference on Advanced Innovation in Engineering and Technology (NCAIET-2015), Vol. 3, Special Issue 1,ISSN(Online) 2321-2004, ISSN(Print) 2321-5528, April 2015.

[3] Z. Dewa and L. A. Maglaras, "Data Mining and Intrusion Detection Systems", International Journal of Advanced Computer Science and Applications, Vol.7, No.1, 2016.

[4] Meenakshi and Geetika, "Survey on Classification Methods using WEKA", International Journal of Computer Applications, Volume 86-No 18 January 2014.

[5] V. Veeralakshmi and Dr. D. Ramyachitra, "Ripple Down Rule learner (RIDOR) Classsifier for IRIS Dataset", International Journal of Computer Science Engineering (IJCSE), Vol. 4, No. 03, ISSN: 2319-7323, May 2015.

[6] J. Patel and Mr. K. Panchal, "Effective Intrusion Detection System using Data Mining Technique", Journal of Emerging Technologies and Innovative Research (JETIR), Vol.2, Issue 6, ISSN-2349-5162, June 2015.

[7] Prof. S. V. Peterraj and S. P. P. Mary, "Study on Data Mining Suitability for Intrusion Detection System (IDS)", International Journal of Data Mining Techniques and Applications, Vol : 01, Issue : 01, ISSN : 2278-2419, Integrated Intelligent Research (IIR), January-June 2012.

[8] A. Narayan M. and T. J. Parvat, "An Intrusion Detection System, (IDS) with Machine Learning (ML) Model combining Hybrid Classifiers", Journal of Multidisciplinary Engineering Science and Technology (JMEST), Vol. 2, Issue 4, ISSN : 3159-0040, April-2015.

[9] Neethu B,"Adaptive Intrusion Detection Using Machine Learning", IJCSNS International Journal of Computer Science and Network Security, Vol, 13, No.3, March 2013.

[10] P. Amudha, S. Karthik and S.Sivakumari, "Classification Techniques for Intrusion Detection – An Overview", International Journal of Computer Applications, Vol 76, No.16,August 2013.

[11] H. A. Nguyen and D. Choi, "Application of Data Mining to Network Intrusion Detection: Classifier Selection Model", Springer-Verlag Berlin Heidelberg 2008.

[12] Saranya. V and Amsaveni. R, "Classification Techniques Applied for Intrusion Detection", International Journal of Computer Science & Engineering Technology (IJCSET), Vol. 3, No. 4, ISSN : 2229-3345, April 2012.

[13] M. R. Kabir, A. R. Onik and T. Samad, "A Network Intrusion Detection Framework based on Bayesian Network using Wrapper Approach", International Journal of Computer Applications(0975-8887), Vol. 166, No.4, May 2017.

[14] https:/ www.techopedia.com/definition/ 3988/ intrusion-detection-system-ids

[15] Y. Y. Aung and M. M. Min, "An Analysis of Decision Tree Based Intrusion Detection System", the First International Conference on Advanced Information Technologies, ICAIT 2017, November 1-2, 2017.

[16] Upendra and Y. K. Jain, "An Empirical Comparison and Feature Reduction Performance Analysis of Intrusion Detection", International Journal of Control Theory and Computer Modelling (IJCTCM), Vol.2, No.1, January 2012.

[17] Neethu B, "Classification of Intrusion Detection Dataset Using Machine Learning Approaches", International Journal of Electronics and Computer Science Engineering, ISSN-2277-1956, Vol.1, No.3-1044-1051.

[18] K. Umamaheswari and S. Janakirama, "Machine Learning in Network Intrusion Detection

System", APPN Journal of Engineering and Applied Sciences, Vol.11, No.2, ISSN 1819-6608, January 2016.

[19] P. R. Patil, Y. Sharma and M. Kshirasagar, "Performance Analysis of Intrusion Detection Systems Implemented using Hybrid Machine Learning Techniques", International Journal of Computer Applications (0975 – 8887), Vol 133, No.8, January 2016.

[20] S. K. Jonnalagadda and R. P. Reddy I, "A Literature Survey and Comprehensive Study of Intrusion Detection", International Journal of Computer Applications (0975 – 8887), Vol 81, No.16, November 2013