A Robust and Efficient Real Time Network Intrusion Detection System Using Artificial Neural Network In Data Mining

Renuka Devi Thanasekaran

Department of Computer Engineering, Anna university, Chennai
renukathanasekaran@gmail.com

ABSTRACT

Today, intrusion detection is one of the major concern in the task of network administration and security. There is a need to safeguard the networks from known vulnerabilities and at the same time take steps to detect new and unseen, but possible, system abuses by developing more reliable and efficient intrusion detection systems. The system must be accurate in detecting attacks with the minimum number of false alarms(wrong detections).Thus an Artificial Neural Network based NIDS is been developed so that the accuracy at which the intrusions are detected increases. In this network intrusion detection system, by using the concept of ensemble binary classification and multiboosting simultaneously it efficiently detects the attack with the low false alarm rate and even at high network traffic. With the use of the Dynamic multiboosting and the database storage the time taken to detect the attacks has been decreased efficiently. By combining the concepts of the Artificial Neural network and the Data mining technique of classification the drawbacks of the later is overcome.

KEYWORDS

NIDS, Ensemble binary classification, Static and Dynamic Multiboosting, Database.

1. INTRODUCTION

Intrusion Detection System (IDS) has been applied to detect intrusion in the network.
Intrusion Detection technology can be defined as a system that identifies and deals with the malicious use of computer and network resources. In the case of detecting data target, intrusion detecting system can be classified as host-based and network-based

- HOST-BASED IDS: Its data is collected from the records of various host activities, including audit record of operation system, system logs, application programs information, and so on.
- NETWORK-BASED IDS: Its data is mainly collected from the network generic stream going through network segments, such as: Internet packets.

Likewise, Intrusion Detection techniques fall into two categories

- ANOMALY DETECTION: is the attempt to identify malicious traffic based on deviations from established normal network traffic patterns.
- MISUSE DETECTION: is the ability to identify intrusions based on a known pattern for the malicious activity.

DOI : 10.5121/ijitcs.2011.1402

While designing an Network intrusion detection system (NIDS), there is a need to correctly detects those attacks and in a short amount of time. So data mining is an most advanced and efficient techniques that could used to design an Data mining refers to the process of extracting descriptive models from large stores of data.

Data mining techniques are being vastly used very recently in various fields. It is taken beneficial steps towards solution of various problems and hence use data mining for solving the problem of network intrusion because of following reasons

- It can process large amount of data.
- User's subjective evaluation is not necessary, and it is more suitable to discover the ignored and hidden information.
- It is a supervised learning methodology.
- It possible to easily perform data summarization and visualization that help the security analysis in various areas.

The recent rapid development in data mining has made available a wide variety of algorithms, drawn from the fields of statistics, pattern recognition, machine learning, Classification methodology in mining maps a data item into one of several predefined categories. These algorithms normally output "classifiers", for example, in the form of decision trees or rules. Hence it would an ideal application in intrusion detection. By gathering sufficient "normal" and "abnormal" data for a user or a program, then apply a classification algorithm to learn a classifier that can label or predict new unseen data as belonging to the normal class or abnormal class.

This paper is organized in the following way. Section 2 briefs about the related mechanisms about the data mining used in the network intrusion detection. Section 3 gives in detail about the architecture of the system being proposed and the mechanism involved in the system. It also contains the results that were obtained when various experiments were conducted on the system.

2. RELATED WORK

Ektefa et al [1] has in the paper proved that Classification decision tree algorithm detects attacks at a very much greater rate than the Service Vector machines(SVM's).Here the same data set were evaluated with the two Data mining approaches. The co-relation between the samples were measured by using the min-max normalization. Results has proved that C4.5 Classification decision tree algorithm is the best by giving very few false alarm rate in the detection of the data set.

Sheen et al [2] has proposed the method of analysis for the best feature selection method for the Network intrusion detection model. In this paper he has analyzed three measures namely Chi-square, Information gain and the Gini index method for feature selection. These are the various filter based approaches that have been used. Among these filter based approaches given upon the open source Windows version 3.4 these three filter approaches were tested. Results have proved that the Information gain when used for the feature selection produces the accurate results by accurately detecting the least prominent attack in the data set.

Chebrolu et al [3] have proposed that the Current intrusion detection systems (IDS) examine all data features to detect intrusion or misuse patterns. Some of the features may be redundant or contribute little (if anything) to the detection process. This paper has identified important input

features in building an IDS that is computationally efficient and effective. The performance of two feature selection algorithms involving Bayesian networks and Classification and Regression Trees is investigated in this paper. The results indicate that significant input feature selection is important to design an IDS that is lightweight, efficient and effective for real world detection systems. Finally, they have proposed an hybrid architecture for combining different feature selection algorithms for real world intrusion detection.

Rokach et al [4] proposed the basic model for the computation of the ensemble based classifier.In this paper they have discussed the various basic blocks that are necessary for the computation of the ensemble classifiers. The core principle is to weigh several individual pattern classifiers, and combine them in order to reach a classification that is better than the one obtained by each of them separately.In this paper they have proposed the best methodology for usinf the ensemble classification approach.

Pedrajas*,et al [5]*discussed the problem of constructing ensembles of classifiers from the point of view of instance selection. Feature selection is aimed at obtaining a subset of the features available for training capable of achieving, at least, the same performance as the whole training set. In this way, feature selection method try to keep the performance of the classifiers while reducing the number of features in the training set. This paper has proposed an algorithm for the ensemble method combining the classifiers generated by each different trained class of the data set.

Weiming Hu et al [6] propose a network intrusion detection system using the Adaboost algorithm. Decision stumps are used as weak classifiers. By combining the weak classifiers for continuous features and the weak classifiers for categorical features into a strong classifier, the relations between these two different types of features are handled naturally, without any forced conversions between continuous and categorical features. The disadvantage of this method is that it does not choose the most contributing feature in the detection. So they have proposed the method an improvement method of including the concept of wagging the algorithm.

Ahmad et al [7] has done a comparative study of the various technique that that are available for the network intrusion detection system using the data mining techniques. This paper they have used Analytic Hierarchy Process for the evaluation of these techniques. As the results suggests the network intrusion detection system involving the concept of the Artificial Neural networks along with the concept of data mining has been proved the best with the maximum efficiency

Webb et al [8] proposed the method of MultiBoosting which is an extension of AdaBoost technique for forming decision committees. It is mainly used to harness both AdaBoost's high bias and variance reduction with wagging's superior variance reduction. In this paper they have proposed the Multi-boosting algorithm.
Thus the system that is being proposed eliminates all the drawbacks of the existing systems. The new methodology of Dynamic multiboosting is being used so that the real time intrusion detection system is much quicker in detecting the attacks. By having creating the log of the various attacks in the database and implementing the multi-threading programming it enables the much efficient and the much faster detection rate in the terms of time involved.

3. PROPOSED SYSTEM

3.1 Packet Capturing

The network intrusion detection system involves the capturing of the packet in real time and with those packets that have been received the intruder in identified. For this purpose the packet

capturing tool called the Winpcap is used. This tool captures the packet that is received from the network and converts those packets to the user reachable form and these packets are given to the

detection process. Normally the various packet capturing tools that are available does not provide much efficiency compared with the Winpcap tool. This tool is the most efficient tool that would be worked under the Windows operating system and it is capable of efficiently capturing packets from other operating systems too.

Since the programming language that is used to develop this IDS is java there needs an connection that is to be establish between the packet capturing tool and the Java. This enables that the captured packets could be accessed by the programming language for the processing. In order to ensure the java connectivity the connectivity module has been designed and implemented.

Various types of analyzers has been designed and each of the Analyzer are designed to handle each type of the protocol in which it has been sent. These analyzer picks up the user defined Characteristics of the protocol and transfers the desired features for the classification.

The layers analyze the packets at the different layers of the OSI model. Each type of the protocol is defined to work under these specific layers

3.2. Ensemble binary classification using C 4.*5*

In the ensemble method of the classification it uses the c4.5 decision tree algorithm. But the normal decision tree algorithm generates the tree in the following procedure.

C4.5 builds decision trees from a set of training data , using the concept of information entropy. The training data is a set $S = s_1,s_2,...$ of already classified samples. Each sample $s_i = x_1,x_2,...$ is a vector where $x_1,x_2,...$ represent attributes or features of the sample. The training data is augmented with a vector $C = c_1,c_2,...$ where $c_1,c_2,...$ represent the class to which each sample belongs.

At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub lists.

This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

The major drawback of this method is that it splits the dataset based on completely the attribute values. This disadvantage is overcome by the usage of the binary C4.5 Algorithm.

Binary C4.5 algorithm works based on the Analyzing the dataset once and then splitting the dataset based on the various class labels involved and then training those instances that belong to that particular class label. By doing this method, the overall accuracy of the classifier that is generated for the particular class label is increased, since the various instances are trained for

that particular class variable. In this method the 5 class variables are trained for their particular instances. And the classifier for the each class label is obtained.

Before generating these class variables we have to do the procedure called the feature selection. From all the variables that are available the attribute that would most efficiently classify the instances belonging to a particular class is generated. Based on these attributes that are selected the classifier is trained.

None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class

3.3. Multi-Boosting

Multi-Boosting is an Artificial Neural network technique where the errors that occurs in the classification process done by the base learning C4.5 algorithm are back propagated so that the final classification that is obtained is the most accurate one.

MultiBoosting is an extension to the AdaBoost technique for forming decision committees. It is a combination of AdaBoost with wagging methods. It is much efficient to get both AdaBoost's high bias and variance reduction with wagging's superior variance reduction. By Using C4.5 as the base learning algorithm, it produce decision committees with lower error than either AdaBoost or wagging significantly more .It offers the further advantage over AdaBoost of suiting parallel execution.

In the real time intrusion detection system there is need to analyze the packets that are coming in the network. **Static multiboosting** is provided by the above said method. Along with the ensemble binary classification this boosting methodology makes the training of those dataset in a more accurate way.

Since in real time classification cannot be done for each and every packet that is received, Dynamic multiboosting is employed. **Dyanamic multi-boosting** is the method by which the database is created and those packets which are classified as the attacks are stored. So that in the real time, the IDS first checks the database for the packets. If an match is found it issues an attack alert. If an match is not found classification is done for the packet.

3.4. Methodology

Network intrusion is detected by the following steps,

1. Train the dataset samples using the ensemble binary classification and multiboosting algorithm.
2. Start the packet capturing tool, and give these packets for classification.
3. Store the packets classified as attack in an database.
4. For the successive packets received in the real time, Check whether those packets were classified as attack in the database.
5. If an match is found display the attack, if not found classify the packet.

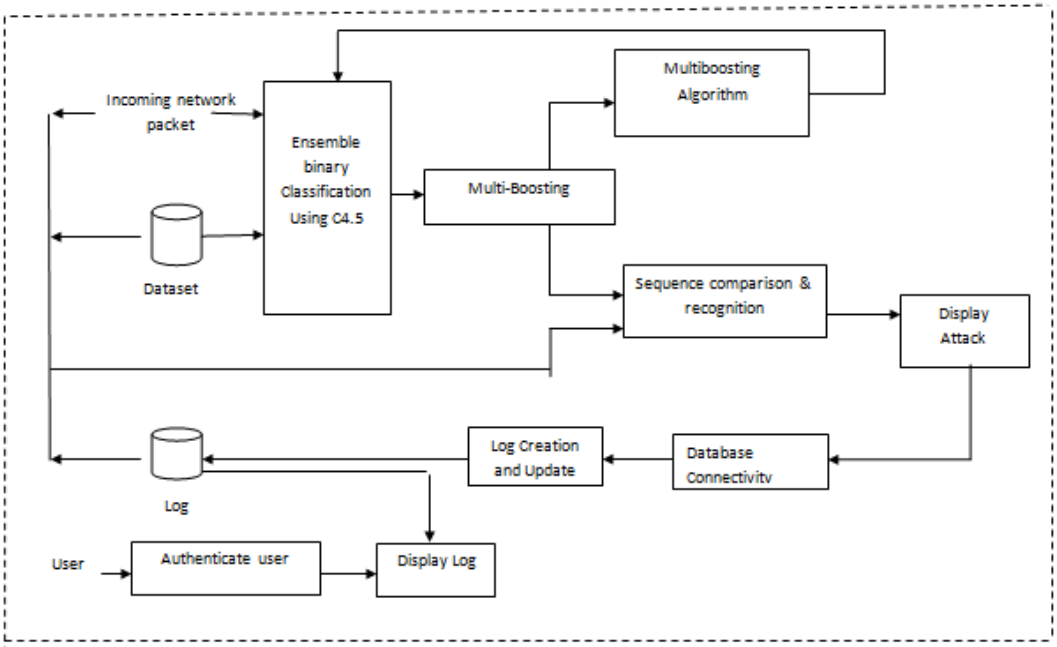

Figure 1

Network Intrusion Detection Mechanism

This database created would act as an log, so that the authenticated user could view the alert details. By creating an web services it is enabled that the authenticated user could see the log from any where in the world.

By the overall process that is shown in the figure 1, the proposed NIDS overcomes the drawback of the time taken to detect the intrusions because of the use of the dynamic multi-boosting technique.

The detection accuracy has increases due to the multi-boosting algorithm as it back propagates the errors. Another advantage of the system is that the use of the supervised learning algorithm so that any errors in the process could be easily identified.

3.5. Motivating Example

For our experiments, we use KDD '99 intrusion data set .In our experiments, we use 494,020 data set for training and 311,029 as test instances. Each record in represents a connection between two IP addresses, starting and ending at some well defined times with a well-defined protocol. Each record is represented using 24 features and it is labeled using the following four attack categories namely Probe, DoS, U2R and R2L under which it contains 24 different kinds of attack. Table 1 gives the number of instances for each group of attack in the data set.

Database creation:

Sqlyog is used to create a database. This database contains the log of the various attacks that have been encountered in the network. The reason for using the Sqlyog is that it has very high speed in connecting with the java console and has high speed in the data retrieval. Multiple threads mechanism could be used to access the various packet sequences that are stored in the database

so that it decreases the time needed for the packet sequence comparison and matching. This increases the overall classification speed of the Network intrusion detection system.

Web Service:

Web Service is been created using java and by impleementing those in the tomcat server only the authenticated user can gain access to the log of the detected attacks and create an responce mechanism to those attacks.These logs are displayed from the log that has been stored in the Sqlyog database.The connecting mechanism has to be included so that the NIDS performs efficiently during the real time implementation.

3.6 Results

From the various experiments conducted, the following results were obtained.

Detection Rate:

Detection rate is an important parameter for evaluation of an NIDS. It is the percentage of detected attacks amongst all attacks. It is calculated using the true positive(TP) and True negative rate(TN).

Detection Rate=(TP/TP+TN)* 100

From our experiments the result obtained in given in the Figure 2

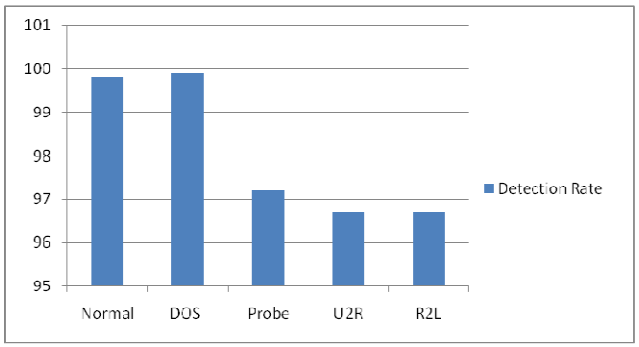

Figure 2:Detection Rate of all types of attack

The reason for such high detection rate of an attack is that the high precision values. This rate is much low for U2R and R2L because of the less number of training samples for these types of an attack.

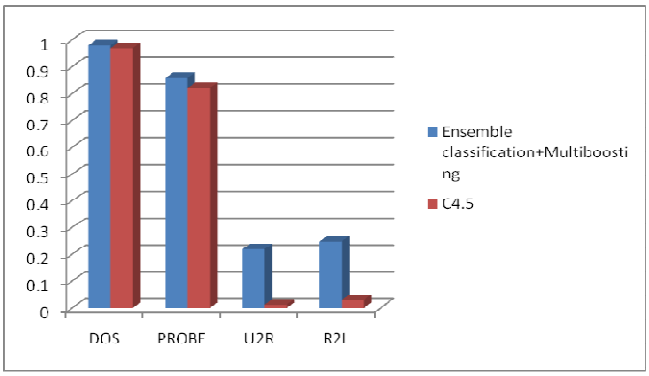

Figure 3
Precision Comparision

From the Fig 3 , it is seen that the ensemble binary classification along with the Multiboosting provides better accuracy than the C4.5 decision tree algorithm as the multiboosting algorithm back propagates the error and with ensemble classification more specific features are analyzed for each type of the attack .

False Alarm Rate:

Another metrics used for estimation is the False alarm rate. It is the normal type being classified as an attack. It is calculated as

False Alarm Rate= (FP/(FP+TN))*100

From our experiments the False Alarm rate is 0.0812.

Time taken:

The speed at which the packet classification is done is calculated real time and the results are given below.

Real time training of the Dataset: .46 sec
Classification time if a packet in real time: 19 milliseconds

From Fig 4 it is seen that since we have used a meta learning algorithm, as the number of packets classified increased the time taken to classify the network packets decreases.

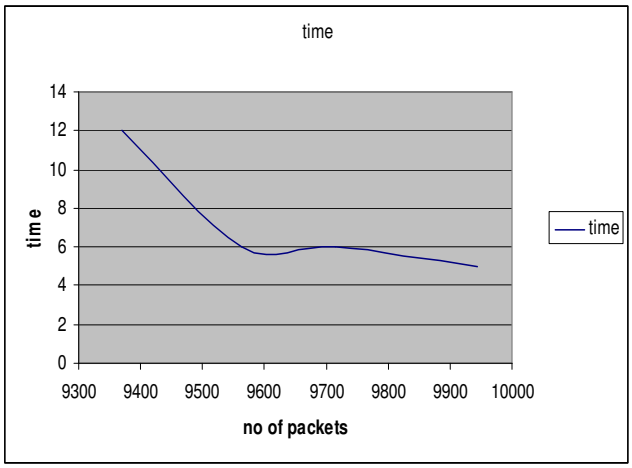

Figure 4
Detection time

The time taken to detect the attacks decreases as to the fact that because of the creating of the log database the need for to classify each of the network packet is eliminated. Thus if any attack is repeated it could be easily detected using the log created.

4. CONCLUSION

From the experiments that were conducted it is seen that the along with the presence of the database and the dynamic multi-boosting the time taken to detect the attacks has increased. By using the data mining methodology of ensemble binary classification using multi-boosting has increased the efficiency and the accuracy at which the attacks have been detected.

REFERENCES

[1] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi, Lilly Suriani Affendey," **Intrusion Detection Using Data Mining Techniques**" International Conference on Information Retriveal and Knowledge Management(CAMP) 2010.

[2] Shina Sheen, R Rajesh," **Network Intrusion Detection using Feature Selection and Decision tree classifier**", TENCON IEEE conference 2008.

[3] S. Chebrolu, A. Abraham, and J. P. Thomas, "**Feature deduction and ensemble design of intrusion detection systems**", Computer & Security, Vol. 24, Issue 4, June 2005, pp. 295-307.

[4] Lior Rokach,"**Ensemble-based classifiers**", Springer Link Science+Business Media B.V. November 2009.

[5] Nicolás García-Pedrajas,"**Constructing Ensembles of Classifiers by Means of Weighted Instance Selection**" IEEE Transactions on neural networks,vol 20,no 2,Feb 2009.

[6] Weiming Hu, Wei Hu, and Steve Maybank," **AdaBoost-Based Algorithm for Network Intrusion Detection**" IEEE transaction on systems,man,and cybernetics-partB: Cybetnetics,vol 38,no 2,April 2008.

[7] Iftikhar Ahmad, Azween B Abdullah, Abdullah S Alghamdi " **Comparative Analysis of Intrusion Detection Approaches**", 2010 12th International Conference on Computer

[8] G. I. Webb, "**Multiboosting: a technique for combining boosting and wagging**", Machine Learning, Vol. 40, 2000, pp. 159-196.

[9] Lui Hui, CAO yonghui "**Research Intrusion Detection Techniques from the Perspective of Machine Learning",** Second International Conference on Multimedia and Information Technology,2010

[10] Naeeam Seilya, Taghi M. Khoshgoftaar, " **Active Learning with Neural Networks for Intrusion Detection**" Knowledge Discovery and Data Mining, 2010.WKDD '10. 3rd International Conference on, Jan. 2010, pp. 601–604.

[11] Thanvarat komviriyavat, Phirivit Snagatanease, "**Network intrusion detection and classification using decision tree and rule based approach**", International conference on machine learning models, technologies and applications, Jan 2009

[12] Juan Wang, Qiren Yang, Dasen Ren," **An intrusion detection algorithm based on decision tree technology"** Asia-Pacific Conference on Information Processing,2009