Proceedings of the International Conference , "Computational Systems for Health & Sustainability"
17-18, April, 2015 - by R.V.College of Engineering,
Bangalore,Karnataka,PIN-560059,INDIA

# Network Attack Detection Using Machine Learning Approach

**Rashmi Hebbar**
M.Tech Student
Computer Science & Engineering
Srinivas Institute of Technology,
Mangalore, Karnataka, India

**Mohan K**
Associate Professor
Department of Computer Science
Srinivas Institute of Technology,
Mangalore, Karnataka, India

**Abstract— With the massive growth of computer networks and the enormous increase in the number of applications that rely on it, network security is becoming very important. Moreover, almost all computer systems in any organization suffer from security vulnerabilities which are both technically difficult and economically expensive to be solved by the manufacturers. Network intrusion Detection System is one of the fundamental components to monitor and analyze the traffic to find out any possible attacks in the network. They are the safety measurements of any network. NIDS plays an important role in privacy security. But the problem is that at what level these NIDS will efficiently able to work? In this paper, the framework for the network intrusion using anomaly method by considering machine learning algorithm is proposed. And the comparison result of using different classifier is achieved.**

**Keywords— Network Intrusion Detection; Machine Learning; Decision Tree; Naïve Bayes; KNN;**

## INTRODUCTION

With the wide spread use of the internet and computer technology, the computer or system violations are increasing at fast rate. Such malicious activities can cause million of damages to the organization. The complete packet inspection is required to examine the data part along with the header content of the packet. Identifying the threat or vulnerabilities in the network itself may reduce the loss in the system. Intrusion detection system (IDS) is security tools that collect information from a variety of network sources, and analyze the information for signs of network intrusions. Generally, network intrusion detection system within the network, monitors the incoming traffic to and from all devices. There are two basic approaches for the intrusion detection technique, i.e Anomaly Detection and Misuse Detection (Signature based ID)[1]. Anomaly Detection is system for detecting the computer intrusion with the set of well defined rules that describes the intrusions. If the captured network signature is not matched with the pre-defined rule, then it is considered as attack. It involves the collection of data relating to the behavior of legitimate users over period of time, and then applies statistical tests to the observed behavior, which determines whether that behavior is legitimate or not. The signature based detection takes the captured packet and known attack rule and produce the derived attack rule as an output. It is mainly used for commercial intrusion detection system.

In this paper, we present the comparative approach of machine learning algorithm have been introduced for the detection of anomalies. As data streams travel across the network, the sniffer captures each packet and constantly decodes and analyzes its content according to the appropriate specification. We capture the real traffic from the wired or wireless medium and perform the intrusion detection based on anomaly and signature based detection. The main aim is to improve the detection accuracy while minimizing the false positive rate [2] and compare their relative performances of using machine learning algorithm.

The rest of the paper is organized as follows: Section II will discuss the overview of Intrusion Detection in the Network. Related works are discussed in section III, section IV says the Methodological approach that includes the theoretical background and proposed system architecture. The experimental setup and result analysis is given in section V and finally section VI concludes the work with the references at the end.

## INTRUSION DETECTION OVERVIEW

When data transfers across the networks, the data passes from the highest layer through intermediate layers to the lowest layer. The lowest layer sends the accumulated data to its destination through the physical network. All network packets passing a certain observation point such as a router are captured without any loss of information and given to the IDS for the analysis of each captured packet.

In general, IDS make use of misuse and anomaly based method to identifying the intrusions. Misuse detection methods are effective for detecting the known attack but fail to detect new attacks whose pattern is not stored in database. Snort [3] is the main tool to identify the known attack signatures and also open source and platform independent tool. Anomaly based method identify the intrusion by analyzing the anomalous behavior from normal profile of data. It has high detection rate for new attack but it produces the false alert.

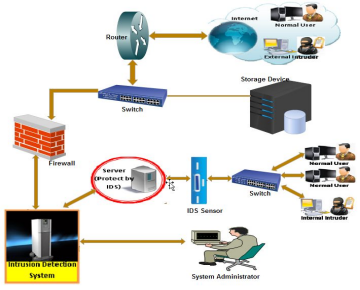


***Fig. 1 Network Intrusion Detection Scenario***

As shown in the Fig. 1 the users request the instances of offered service through the internet. The internal or external attacker in the network tries to gain the information about the legitimate user and harm the system. IDS should analyze the information gathered by the sensors, and return a fusion of the input of the sensors to system administrator or

intrusion prevention system. System administrator carries out the prescriptions given by the IDS and has detailed information about the system.

## RELATED WORK

Intrusion Detection technique started in early 80's and number of the technique has been introduced. One such technique of identifying the network intrusion is through machine learning techniques. This technique has ability of learning and improving its performance over time. Focus on constructing a system which can optimize the performance in a loop and can exchange its execution strategy according to the feedback. In 2007, M.Panda and M.R. Patra [5] proposed a method using Naïve bayes to detect the specific attack. The evaluation of this method is done on KDD 99 dataset. Roshni Dubey and P.Nandan Pathak [6] used KNN based classifier to cluster and analyze the intrusion. They have successfully able to perform the result by feature reduction and cluster based method. Gary Stein [7] applied genetic algorithm and Decision tree for intrusion detection. Mainly the feature reduction is done based on genetic algorithm. In 2010, H.Nguyen et.al [8] applied c4.5 and Bayesnet for intrusion detection. Mohammdreza Ektefa et.al [9] proposed intrusion detection system based on c4.5 and SVM. They revealed the results showing that c4.5 algorithm has better detection rate compared to its predecessor.

In this paper, to improve the detection accuracy, we combine three classifiers (Decision tree, Naïve Bayes and KNN) rather than using them individually. The comparative graph shows the ability of detection with each proposed method.

## METHODOLOGICAL APPROACH

### *A.* Background

Decision tree technology is a common, intuitionist and fast classification method [11]. The decision tree classifier is one of the good approaches to multistage decision making. The basic idea involved in decision tree is to break up a complex decision into a union of several simpler decisions, hoping the final solution obtained would resemble the intended desired solution. Here each node represents a feature name, each leaf indicates a class label and each branch represents an outcome of the associated node. It can classify a large amount of data with faster learning speed. One of the advantages of decision tree is that it does not require users to have a lot of background knowledge. In the process constructing decision tree, selection of testing attributes and how to divide sample set is very crucial. Different Decision tree algorithm uses different technologies. At present there are many decision algorithms present; such as ID3, SLIQ, CART, and CHAID and so on.

***Advantages:***

- Easy to understand, simple and fast approach.
- Leads to a good accuracy (may depend upon the data).
- It supports an incremental approach.

Naive Bayes classifier [10] is a simple probabilistic classifier and results in reducing the false positive alert rate. This is based on Baye's theorem with strong assumptions. It can be used to predict the class label like normal or intrusion of the given network traffic by calculating the probability. This technique is generally used for intrusion detection in combination with statistical schemes. The naïve Bayesian (NB) algorithm is used for learning task, where a training set with target class is provided. The naïve Bayesian (NB) algorithm is used for learning task, where a training set with target class is provided. Training set is described by attributed A1 through An, and each attribute is described by attribute values $a1$, $a2$ … $an$ associated with class C. The objective is to classify an unseen example, whose class value is unknown but attribute values are known. The Bayesian approach to classifying the unseen example is to assign the most probable target class.

***Advantages:***

- Construction is easy and also takes short computational time.
- It can be applied to large dataset since it does not involve in complicated parameter.
- Interpretation of knowledge representation.

k-NN is one of instance-based learning method [12] where the function is only rounded off locally and all computation is deferred until classification. Nearest neighbor search is one of the most popular learning and classification techniques introduced by Fix and Hodges. The k-NN algorithm is the simplest algorithm of all machine learning algorithms. It does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbors of each point. The principle is to find a predefined number of training samples closest in distance to the new point, and predict the class label from these. The number of samples can be a user-defined constant (k-nearest neighbor learning), or vary based on the local density of points.

***Advantages:***

- Understanding is easy and simple implementation steps.
- Lazy learning methods are faster at training time.

### *B.* *Proposed System*

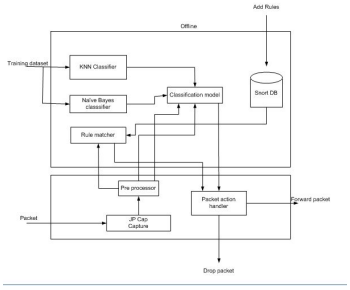


***Fig. 2 Architecture of Proposed Method***


 

Many experiments have been conducted to achieve the final results. The architecture of proposed method is shown in the Fig. 2. Initially KDD cup 99 [13] is used for the training purpose for intrusion detection. This dataset is one of the most realistic publicly available sets that include actual attacks. The KDD dataset was acquired from raw tcpdump data and this dataset includes three independent sets: "whole KDD", "10% KDD", and "Corrected KDD".

Once the training is done the classifier module either decision tree, Naïve bayes or KNN method is used to identify the true intrusion from the incoming network traffic. Winpcap is used as the real time packet capturing tool. From the host machine, it captures the in-bound and out-bound network traffic (packets) for auditing. If any packet is matched with the predefined classifier, alarm will be generated for those abnormal packets and the central log of malicious packet can be stored for the further and detailed analysis.

The main task is to compare the result of using the different classifier and how well they can be response to each captured traffic with the minimizing false positive.

## EXPERIMENTAL EVALUATION

### *C. Experimental Setup*

We have done the whole experiment in the virtual machine based environment. The different guest os acts as the client. The traffic from the host machine is allowed on each client for the intrusion detection. We install the winpcap for capturing the real network traffic. For testing feasibility of different classifier in proposed method, we have used KDD intrusion dataset as a training dataset for all classifier. KDD 10% contains the 4,94,021 network records with 41 features and 24 different classes. All 41 feature may decreases the detection accuracy and speed [14]. So we are using the subset of the features from KDD dataset and they are relevant to each type of attacks. Non-relevant features may effect the overall detection accuracy. The table 1 summerizes the 11 features with the better gain.

***Table 1. Gain of the 11 features in KDD Dataset***

| Feature No. | Feature Name | Gain |
|---|---|---|
| 2 | protocol_type | 0.3024 |
| 3 | service | 0.5709 |
| 5 | src_bytes | 0.6460 |
| 6 | dst_bytes | 0.5383 |
| 23 | Count | 0.6193 |
| 35 | dst_host_diff_srv_rate | 0.3013 |
| 36 | dst_host_same_src_port_rate | 0.3847 |
| 39 | dst_host_srv_serror_rate | 0.0801 |
| 37 | dst_host_srv_diff_host_rate | 0.0681 |
| 41 | dst_host_srv_rerror_rate | 0.0894 |
| 40 | dst_host_rerror_rate | 0.0456 |

Once the training process is done, the weka classifier automatically pre-processing the data to remove the less correlation data with intrusion detection. The real network traffic is captured and analyzed and given for the different classifier to find any intrusion.

The most important factor involved in the performance evaluation of IDS is listed below.

True Positive Rate (TPR) : Also called as Detection Rate (DR). It is the probability of correctly detected intrusion record.

$$DR = \frac{total\_detected\_attacks * 100}{total\_attacks}$$

False Positive Rate (FPR) : The probability of intrusion alert, when there is no intrusion. For the efficient system FPR must be minimum.

$$FPR = \frac{total\_misclasified\_process * 100}{total\_normal\_process}$$

### *D.* Result Analysis

The bellow table shows the performances of three different classifier based on correctly classified instances and incorrectly classified instances, Kappa statistics, Mean absolute error, Root mean squared error and relative absolute error and the time taken to build the models. The comparison is done on the 11 attributes of KDD dataset.

***Table 2. Comparison of the result for classifier with 11 attributes***

| Parameter | Classifier | | |
|---|---|---|---|
| | Decision | Naive | KNN |
| Correctly | 96.54% | 98.18% | 99.01% |
| Incorrectly | 3.43% | 1.28% | 0.44% |
| Kappa statistics | 0.930% | 0.991% | 0.878% |
| Mean absolute | 0.378% | 0.152% | 0.11% |
| Root mean | 0.175% | 0.339% | 0.018% |
| Relative absolute | 11.60% | 19.65% | 14.281% |
| Root relative | 35.07% | 24.84% | 13.45% |

Now we compare the result of hybrid method. Initial comparison is done on the classifier i,e Naïve bayes, Decision Tree and KNN. We compare the result by capturing the real time network traffic and conclude that which algorithm is best suited for the intrusion detection.

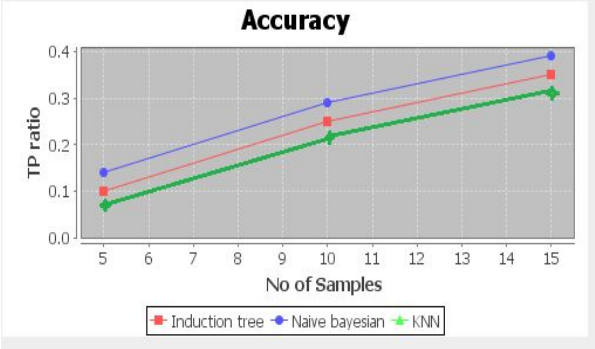


***Fig. 3 TP Comparison of Proposed Method***

The Fig. 3 shows the True positive rate of the Naïve Bayes, Decision tree and KNN algorithm. For the good IDS True positive should be high. Above figure shows that TPR of Naive bayes algorithm is higher for the reduced attribute set.

Fig. 4 shows the false positive rate of Naïve Bayes, Decision Tree, and KNN algorithm when we run with the 11 attributes of dataset.

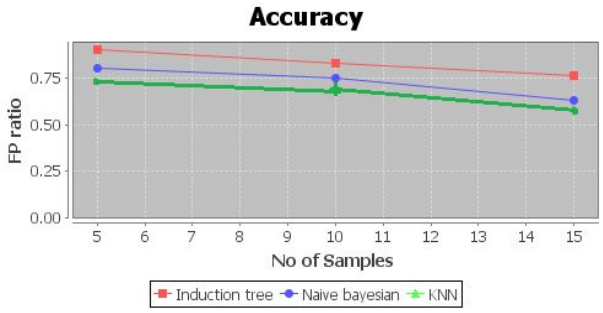


***Fig. 4 FP Comparison of Proposed Method***

For good IDS false positive rate should be low. As we absorb from above fig. FPR of naïve bayes algorithm is lower compared to other approaches. From Fig. 3 and 4 it is clear that accuracy, FPR and TPR of naïve bayes algorithm is better.

In general, proposed hybrid method is capable of detecting higher number of intrusion with low false alerts and with high accuracy. We are successfully able to get better result by combining the different methodologies. Moreover, proposed hybrid method has low computational and communicational overhead compare to the other classifier like Artificial Neural Network (ANN) [15] of machine learning approach. The use of multiple classifiers successfully improves the intrusion detection.

## CONCLUSION

In this paper, we propose the security framework that integrates hybrid NIDS. The main purpose of using hybrid approach is to improve the efficiency when compared to the single approach. We have used both signature based and anomaly based detection method to improve the detection accuracy and also comparison made on different machine learning algorithm and achieved a good result. Thus the proposed method looks very promising and ensures the better feasibility of securing the any system.

## REFERENCES

[1]. Prerika Agarwal, S.Satapathy, "Implementation of Signature-based Detection System using Snort in windows", International Journal of Innovation & Advancement in Computer Science (IJIACS), May 2014.

[2]. C.Modi, Dhiran Patel, "A Novel hybrid- Network Intrusion Detection System in Cloud Computing", IEEE, 2013.

[3]. R. Vanathi, S. Gunasekaran, "Comparision of Network Intrusion Detection System in Cloud Computing Environment", International Conference on Computer Communication and Informatics (ICCCI -2012), Jan. 10 – 12, 2012, Coimbatore, INDIA

[4]. J. Allen, A. Christie, W. Fithen, J. McHugh, and J. Pickel,"State of the practice of intrusion detection technologies," in CMU/SEI-99-TR-028, 2000.

[5]. M. Panda and M.R Patra, "Network Intrusion Detection using naïve bayes", International Journal of Computer science and Network Security (IJCSNS) Volume-7, No 12, December 2007, pp-258-236.

[6]. R. Dubey, P. Nandan Pathak, " KNN based Clasiffier system for Intrusion Detection", International Journal of Advanced Computer Technologies(IJACT), Volume 2, N0 4, ISSN: 2319-7900.

[7]. G. Stein, B.Chen, "Decision Tree Classifier for Network Intrusion Detection with GA based feature selection", University of Central Florida. ACM-SE 43, Proceeding of 43rd annual Southeast regional Conference, Volume-2, 2005 ACM, New York, USA.

[8]. H. Nguyen, K. Franke, S. Petrovi'c, " Improving effectiveness of Intrusion Detection by Correlation Feature Selectio", International Conference on Availability, Reliability and Security, pp. 17-24. IEEE 2010.

[9]. M. Ektefa, S. Memar, et.al. "Intrusion Detection using Data Mining Technique", Proceedings of IEEE International Conference on Information Retrieval & Knowledge Management, Exploring Invisible World, CAMP'10, 2010, pp200-203.

[10]. A. Jain, S.Sharma, M. S. Sisodia, "Network Intrusion Detection by using Supervised and Unsupervised Machine Learning Techniques: A survey", International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 1, Issue 3, Nov 2011.

[11]. Y. K. Jain and Upendra, " An Efficient Intrusion Detection Based on Decision Tree Classifir Feature Reduction", International Journal of Scientific and Research Publications, Volume 2, Issue 1, January 2012.

[12]. http://scikit-learn.org/stable/modules/neighbors.html

[13]. KDD cup 1999 [Online] Available: http://kdd.ics.uci.edu/databases/kddcup99/kdcup99.htm

[14]. S.S Sathya, R.G. Ramani, and K. Sivaselvi, "Discriminant Analysis based Feature Selection in KDD Intrusion Dataset", International Journal of Computer application, vol 31, n0-11, pp.1-7,2011.

[15]. J. Singh, M. J. Nene, "A Survey on Machine Learning Techniques for Intrusion Detection Systems", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 11, November 2013.


 