Data Mining Techniques for Intrusion Detection

Prashant Joshi
M. Tech (CSE) Scholar
Uttaranchal University

Kamlesh Padaliya
Assistant Professor
GEHU, Bhimtal Campus

Dr.Anchit Bijalwani
Assistant Professor
Uttaranchal University

***Abstract*— Intrusion detection can be defined as an act of detecting actions that attempt to compromise the confidentiality, integrity or availability of any network resource. In this paper we discuss the different data mining techniques for intrusion detection. We review some of the existing ensemble methods used in intrusion detection. We also propose an ensemble method for the problem of intrusion detection.**

***Keywords*— Intrusion detection, ensemble method**

I. INTRODUCTION

With the availability of low cost powerful computers coupled with the growth of the Internet and high-speed networks, security has become a matter of immense concern. There is a potential threat to the information stored in our system from attackers. These attacks can be called Intrusions, which are any set of actions that threaten the integrity, availability or confidentiality of a network resource.

By integrity, it means that exact data should arrive at receiver's end. By availability, it means that the service should be available whenever required while confidentiality means that the data should only be seen/accessed only by authorized users. There are several attacks corresponding to integrity, availability and confidentiality of a network resource. They are as follows.

Threat to Integrity: Modification, Masquerading, Replaying, Non Repudiation

Threat to Availability: Denial of Service

Threat to Confidentiality: Snooping, Traffic Analysis

There are three main phases in network security. Intrusion detection aims at defining techniques which allow detecting attacks while they are being performed. Intrusion prevention aims at defining strategies and policies which can prevent intrusion from occurring or reduce the probability of such events. Intrusion reaction involves forensic analysis and other such activities.

This paper focuses on Intrusion Detection Systems (IDS). Part II presents an overview of different detection methods. Part III covers various data mining techniques involved in intrusion detection. In part IV, we discuss the experiment carried out on network security dataset using different machine learning algorithms.

II. INTRUSION DETECTION

Intrusion detection systems (IDSs) are monitoring devices that can detect any malicious activity. The goal of an Intrusion Detection System (IDS) is to detect malicious traffic. To accomplish this, the IDS monitors all incoming and outgoing traffic in the network.

There are basically two types of detection techniques [1]. They are

a) Anomaly based detection: In this technique, the normal behavior of the network is studied. Any deviation from this normal behavior triggers an alert by the IDS.
b) Misuse based detection: It is also known as knowledge based detection. This method makes use of signatures of previously detected attacks in order to detect new attacks. In this method, the IDSs have an access to a database which contains signatures of all previously detected attacks. Signature is pattern or a sequence of instructions which define an attack

Though, misuse based detections simple and accurate, it fails when there is any new type of attack or any new variant of known attack. Thus, IDS is not able to detect them. Other drawbacks of this technique are large number of false positives and false negatives. False positive occurs when a normal activity is mistakenly classified as malicious while false negative occurs when a malicious activity is mistakenly classified as normal.

These limitations have led to an increasing interest in intrusion detection techniques based on data mining.

III. DATA MINING TECHNIQUES FOR INTRUSION DETECTION

Data Mining is the process of discovering interesting and useful patterns and relationships in large volumes of data. Data mining can be used to uncover hidden patterns within large amounts of data and these hidden patterns can potentially be used to predict future behavior. The goal of data mining is to extract information from a data set and transform it into an understandable structure for further use.

As discussed in the previous section, there are some cases in which Intrusion Detection Systems are unable to detect the attack on the network. These attacks are generally the ones which are variants of previously seen attacks. For this reason, learning mechanisms must be implemented in Intrusion Detection Systems to detect and prevent these attacks without

having to wait for updates or patches. Machine Learning can be applied for this purpose. Just like machine learning enables a computer to learn how to make predictions it can be used in intrusion detection to predict if an action is normal or malicious. However, it is not very simple to apply machine learning in Intrusion Detection Systems.

Machine learning can be divided into two major classes depending on their learning technique: supervised and unsupervised. There is also another type of learning known as semi supervised learning. Supervised learning is the task of inferring a function from labeled training data, a set of training examples. A supervised learning algorithm analyzes the training data and produces an inferred function that can be used for mapping new examples. This can correctly determine the class labels for unseen instances. While unsupervised learning does not require any training data. Semi-supervised learning is a class of supervised learning tasks and techniques that makes use of both labeled and unlabeled data i.e. it uses a small amount of labeled data and a large amount of unlabeled data.

It is very important to remove the redundant and irreverent attributes from the dataset before it is fed to the machine learning algorithm used as classifier. This is done by feature selection. Feature Selection is a pre- processing step and independent of the machine learning algorithm applied. It is also called subset selection or variable selection. In feature Selection, a subset of features available in the data is chosen to be used in the learning process. It is important as all features of the data are not required in learning process.

In ensemble approach several machine learning algorithms are combined for classification. The idea behind ensemble approach is to exploit the strengths of each algorithm in it and to obtain an efficient and robust classifier [2]. One challenge in using ensemble approach for classification is selection of constituent algorithms of the ensemble and the decision function which combines the results of these algorithms. The two techniques used to combine algorithms in an ensemble are Bagging and Boosting. The algorithms are made to run in parallel in ensembles using Bagging Technique while in ensembles using Boosting technique the algorithms are made to run sequentially.

Ensemble approaches were introduced in the late 80s. Hansen and Salamon [3], in the year 1990, showed that the combination of several Artificial Neural Networks (ANNs) can considerably improve the accuracy of the predictions. A number of papers were written on the subject in the subsequent years. However, ensemble approach was used for the first time for intrusion detection in the year 2003.

Maximizing detection accuracy and minimizing false alarm rate are two major challenges in designing anomaly intrusion detection systems. This issue has been addressed by Zainal, Marrof and Shamsuddin their paper "Ensemble Classifiers for Network Intrusion Detection System" [4].In this paper, they propose an ensemble of one-class classifiers with different learning paradigms. The techniques deployed in this ensemble model are Linear Genetic Programming (LGP), Adaptive Neural Fuzzy Inference System (ANFIS) and Random Forest (RF). The performances of individual classifiers were evaluated and an ensemble rule was formulated. Before classification, feature selection process was also performed to improve the detection process. The results of experiments show an improvement in detection accuracy for all classes of network traffic; Normal, Probe, DoS, U2R and R2L.

Mukkamala, Sung and Abraham [5] showed that an ensemble composed of different types of ANN, Support Vector Machines (SVM) with Radial Basis Function (RBF) kernel and Multivariate Adaptive Regression Splines (MARS) combined with the bagging techniques outperforms approaches using single algorithm.

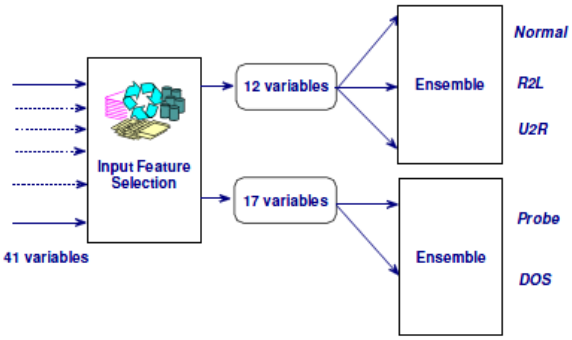


Figure: Ensemble developed by Abraham and Thomas [6]

The above figure shows the ensemble developed by Abraham and Thomas for the problem of intrusion detection. The authors have applied feature selection to reduce the features of the KDD99 dataset from 41 to 12 for the three classes normal, remote to local and user to root. For classes denial of service and probe they have reduced the number of features from 41 to 17. Their model has significantly increased the performance of the IDS.

Giorgio Giacinto and Roli in their paper "Intrusion Detection in Computer Networks by Multiple Classifier Systems" [7] propose pattern recognition approach to network intrusion detection based on the multiple classifier systems. Panda and Patra, in their paper "Ensemble Voting System for Anomaly Based Network Intrusion Detection" [8], analyze the performance of classifiers in heterogeneous environment using voting ensemble system for detecting intrusions using anomaly based technique. Results onKDDCup1999 dataset demonstrate that the voting ensemble technique yield significantly better results in detecting intrusions as compared to other techniques.

The conclusion that can be drawn from all reviewed work is that the ensemble approach generally outperforms traditional approaches in which only one algorithm is used. Thus, an ensemble is a very efficient technique to compensate for the low accuracy of a set of weak learners. Moreover, proper feature selection can further enhance the performance of the ensemble.

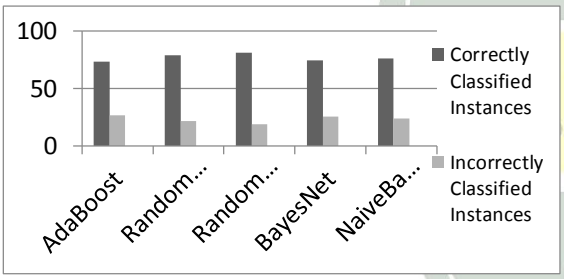
IV. EXPERIMENT

In the experiment we compare the performance of different machine algorithms. The experiments were carried out on an Intel Core 2 Duo Processor 2.20 GHz. RAM with 4GB RAM. The tool used for the experiment is WEKAData Mining System (version 3.6).

The dataset taken is the KDD cup 1999[9] dataset built by the Defense Advanced Research Projects Agency (DARPA) in 1998 during the DARPA98 IDS evaluation program. The data set contains network traffic data. The Data set has millions of records and features labeled from 1 to 41. The attacks identified can be categorized into four main categories Denial of Service (DoS), Remote to Local (R2L), User to Root (U2R) and Probing. Denial of Service attack can be characterized as attacks which prevent or deny authorized users access to requested service or resource. The Remote to Local attack is one in which the attacker gains remote access to any unauthorized node or resource in the network. In User to Root attack, any local normal user can gain access to the administrative rights of the network and act as master node. The probing attack is one in which the attacker keeps a track of the information being shared in the network. The attacker looks for the opened and vulnerable ports which he/she can exploit later.

V. RESULT

The performance of various algorithms on KDD Data Cup '99 has been compared in the following chart.



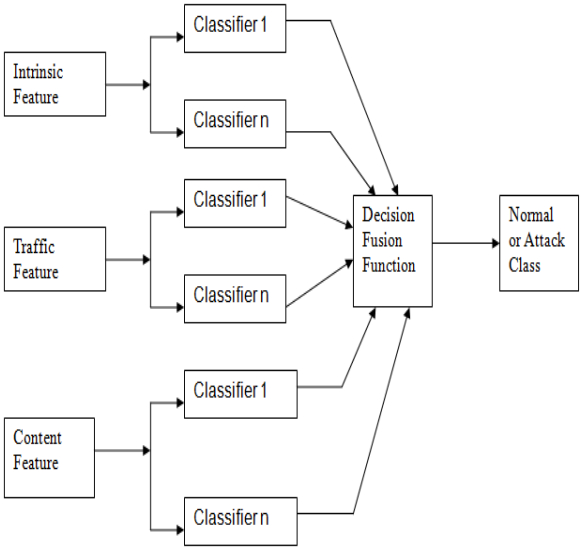
It can be seen that Random Tree has outperformed all other algorithms in correctly classifying the entries of the dataset. On the other hand, AdaBoost has performed most poorly among all algorithms.

VI. FUTURE WORK

The future work includes development of new ensemble method for different network security datasets. The proposed methodology is shown in the following figure.

Intrinsic features includes general information like the duration in seconds of the connection, the protocol type etc. Traffic Feature includes information such as number of connections with the same destination host. Content features include information such as number of failed login attempts or information about payload [10].

We shall be developing an ensemble of multiple classifiers. If single classifier is used for classification, a critical point may be reached from where no further improvement in classification is possible. Applying multiple classifiers will help to push the critical point further



VII. CONCLUSION

In this paper, we have reviewed various data mining techniques that can be used for the detection of intrusions. We have also discussed different ensemble methods for intrusion detection. The performances of different machine learning algorithms on KDD Cup 99 network security dataset. Finally, we a propose an ensemble of multi classifier that can be used to develop a better and more effective Intrusion Detection Systems.

REFERENCES

[1] Herv´eDebar,"An Introduction to Intrusion-Detection Systems"

[2] David Opitz, Richard Maclin,"Popular Ensemble Methods: An Empirical Study"

[3] Lars Kai Hansen and Peter Salamon.Neural Network Ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(10):993{1001, October 1990

[4] AnazidaZainal, MohdAizainiMarrof and SitiMriyamShamsuddin, "Ensemble Classifiers for Network Intrusion Detection System", Journal of Information Assurance and Security 4 (2009), 217-225

[5] SrinivasMukkamala, Andrew H. Sung, and Ajith Abraham. Intrusion detection using an ensemble of intelligent paradigms. Journal of Network and Computer Applications – Special issue on computational intelligence on the internet, 28(2):167{182, April 2005}

[6] Ajith Abraham and Johnson Thomas. Distributed Intrusion Detection Systems: A Computational Intelligence Approach, volume 5, pages 105{135. Idea Group Inc. Publishers, 2005

[7] Giorgio Giacinto and Fabio Roli, "Intrusion Detection in Computer Networks by Multiple Classifier Systems" In Proc. 16th International Conference on Pattern Recognition (ICPR'02)

[8] Mrutyunjaya Panda and ManasRanjanPatra ,"Ensemble voting System for Anomaly Based Network Intrusion Detection", International Journal of Recent Trends in Engineering, Vol 2, No. 5, November 2009

[9] MahbodTavallaee, EbrahimBagheri, Wei Lu, and Ali A. Ghorbani "A Detailed Analysis of the KDD CUP 99 Data Set" Proceedings of the Second IEEE Symposium on Computational Intelligence forSecurity and Defence Applications 2009, 2009-07-10

[10] Cost-Based Modeling for Fraud and Intrusion Detection: Results from the JAM Project, volume 2, Hilton Head, South Carolina, January 2000