Reducing Unwanted Attribute in Intruder File Based On Feature Selection and Feature Reduction Using Id3 Algorithm

Santosh Nikam , Prof.Saurabh Mandloi, Prof. Parmalik Kumar
Computer Science & Engineering Department
PCST Ratibad, Bhopal
santosh18636@gmail.com

ABSTRACT : Reduction and selection of intruder attribute in intrusion detection system play an important role in process of detection. The huge number of attribute in intruder induces a problem in detection process and increase more time in decision making process. In current research trend some authors used some standard technique for feature reduction such as PCA, PCNN and neural network, but these methods not consider all features for processing fixed some number of feature. In this paper proposed a feature selection and feature reduction method based on improved ID3 algorithm. The proposed algorithm select multiple feature for reduction and the reduce feature set participant the process of detection. The reduce feature of network file classified by ID3 classification algorithm. The ID3 algorithm in the case of small data size, if sizes of data are increase the selection of attribute process raised some problem related to feature selection. For the improvement of this problem used RBF function for increasing the biased value of feature and feature subset selection. In this paper tried to propose a very simple and fast feature selection method to eliminate features with no helpful information on them. Result faster learning in process of redundant feature omission. We compared our proposed method with three most successful similarity based feature selection algorithm including Correlation Coefficient, Least Square Regression Error and Maximal Information Compression Index. For the validation and performance evaluation of proposed algorithm used MATLAB software and KDDCUP99 dataset 10%. This dataset contains approx 5 lacks number of instance. The process of result shows that better classification and reduce time instead of another feature reduction.

I. INTRODUCTION

The performance of intrusion detection system depends on classification of unknown types of attacks. The detection of unknown types of attack is very difficult due to large number of attribute and huge amount of network data. For the improvement of unknown attack feature reduction is important area of research. The reduction process reduces the large number of attribute and improved the detection of intrusion detection system. In the process of feature reduction various algorithm are used such algorithm are principle of component analysis and neural network. The reduction process used PCA method this method is static reduction technique, reduces only fixed number of attribute. The fixed number of feature reduction process not justify the value of feature it directly reduces the feature. On the consideration of computational time feature reduction is also an important aspects, the reduces feature increase the processing of detection ratio. Many methods have been proposed in the last decades on the designs of IDSs based on feature reduction technique.With the tremendous growth of network-based services and sensitive information on networks, network security is becoming more and more importance than ever before. Intrusion detection techniques are the last line of defenses against computer attacks behind secure network architecture design, firewalls, and personal screening. Despite the plethora of intrusion prevention techniques available, attacks against

computer systems are still successful. Thus, intrusion detection systems (IDSs) play a vital role in network security. Symantec in a recent report uncovered that the number of fishing attacks targeted at stealing confidential information such as credit card numbers, passwords, and other financial information are on the rise, going from 9 million attacks in June2013 to over 33 millions in less than a year. One solution to this is the use of network intrusion detection systems (NIDS) that detect attacks by observing various network activities. It is therefore crucial that such systems are accurate in identifying attacks, quick to train and generate as few false positives as possible. Internet has rapidly become one of the main communication methods in our society. Various types of internet application and usage are available more and more. Increasing usages of network applications also increase security risks to internet users, to prevent unwanted or dangerous threats.

1.1 INTRUSION DETECTION

An Intrusion Detection System (IDS) inspects the activities in a system for suspicious behavior or patterns that may indicate system attack or misuse. There are two main categories of intrusion detection techniques; Anomaly detection and Misuse detection. The former analyses the information gathered and compares it to a defined baseline of what is seen as "normal" service behavior, so it has the ability to learn how to detect network attacks that are currently unknown. Misuse Detection is based on signatures for known attacks, so it is only as good as the database of attack signatures that it uses for comparison. Misuse detection has low false positive rate, but cannot detect novel attacks. However, anomaly detection can detect unknown attacks, but has high false positive rate.

An intrusion detection system gathers and analyzes information from various areas within a computer or a network to identify possible security breaches. In other words, intrusion detection is the act of detecting actions that attempt to compromise the confidentiality, integrity or availability of a system/network. Traditionally, intrusion detection systems have been classified as a signature detection system, an anomaly detection system or a hybrid/compound detection system. A signature detection system identifies patterns of traffic or application data presumed to be malicious while anomaly detection systems compare activities against a "normal" baseline. On the other hand, a hybrid intrusion detection system combines the techniques of the two approaches. Both signature detection and anomaly detection systems have their share of advantages and drawbacks. The primary advantage of signature detection is that known attacks can be detected fairly reliably with a low false positive rate. The major drawback of the signature detection approach is that such systems typically require a signature to be defined for all of the possible attacks that an attacker may launch against a network. Anomaly detection systems have two major advantages over signature based intrusion detection systems. The first advantage that differentiates anomaly detection systems from signature detection systems is their ability to detect unknown attacks as well as "zero days" attacks. This advantage is because of the ability of anomaly detection systems to model the normal operation of a system/network and detect deviations from them. A second advantage of anomaly detection systems is that the aforementioned profiles of normal activity are customized for every system, application and/or network, and therefore making it very difficult for an attacker to know with certainty what activities it can carry out without getting detected. However, the anomaly detection approach has its share of

drawbacks as well. For example, the intrinsic complexity of the system, the high percentage of false alarms and the associated difficulty of determining which specific event triggered those alarms are some of the many technical challenges that need to be addressed before anomaly detection systems can be widely adopted.

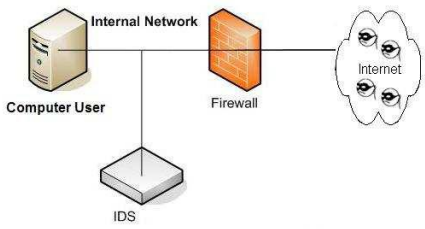

Figure:Intrusion detection systems in the network environment

Types of intrusion detection systems there are two types of intrusion detection systems that employ one or both of the intrusion detection methods outlined above. Host-based systems base their decisions on information obtained from a single host (usually audit trails), while network-based intrusion detection systems obtain data by monitoring the traffic in the network to which the hosts are connected. An intrusion detection system dynamically monitors the events taking place in a monitored system, and decides whether these events are symptomatic of an attack or constitute a legitimate use of the system. Figure depicts the organization of IDS where solid arrows indicate data/control flow while dotted arrows indicate a response to intrusive activities.

II. IDS TECHNIQUES

Feature selection and feature reduction play a vital role in the field of intrusion detection system. The feature reduction improved the performance of intrusion detection process. This reduces feature of intruder file are ideal data, those data are not participant in any attack mode of activity as knows attack and unknown attack. The process of reduction adapts some heuristic function such as genetic algorithm, ANT colony optimization and many more technique for reduction. In this chapter discuss several optimisation and feature reduction technique of intrusion detection. Classification technique is a collection of several methods, which aim to exploit tolerance for in distinctness, uncertainty and incomplete fact to achieve tractability, robustness and low solution, cost. As soft computing techniques can also be used for machine learning, different Classification techniques have been used for intrusion detection system such as Neural network (NN), Support vector machines (SVM), Artificial Neural Networks (ANN), Decision tree (DT), KNN and Clustering and outlier detection. Genetic algorithm (GA) field is one of the upcoming fields in computer network security, especially in intrusion detection systems (IDS).

2.1 ANOMALY DETECTION

As misuse detection was based on previously known patterns anomaly detection may detect also something that has not yet been discovered. It is also worth noting that while intrusion detection assumes all the matching activities as malicious, anomaly detection does not assume all anomalies necessarily malicious. [10] It depends on the environment and the rules and regulations whether the detected anomaly is malicious or not.

Network traffic anomaly detection is based on two presumptions. The first presumption is that network traffic has distinguishable characteristics in normal conditions. A model of these normal conditions can be created with parameters. The second presumption is that deviations from this normal model are rare and potentially might be a result of intrusive activity. These two presumptions are according to what is presented in the field literature.

2.1.1 ANOMALY DETECTION AS A PROCESS

As a process, anomaly detection can be divided into two phases. In the first phase a model of normal network traffic is created. This model can be derived or learned from training data using model generation algorithms or mathematical models. In the second phase traffic is monitored for deviations from the normal model. The model of normal network traffic is created by using features from the traffic. Feature in the context of anomaly detection means a value or symbol which describes the network traffic. These features should represent the traffic behaviour and characteristics but in the same time they should not contain any redundant information in order to be as lightweight as possible.

2.1.2 STATISTICAL BASED ANOMALY DETECTION

In a statistical based method anomalies are detected from statistics. Statistical based methods create models based on history. These models are then compared to the current situation and deviations between these models are considered as anomalies. Once a deviation is monitored its severity is then evaluated and graded. The more severe the anomaly is the higher the grade is. [15] For example, the average number of times a user has accessed the network daily is compared to the current amount. If the current number of access to the network exceed the average number by one or two it is not maybe considered as a severe anomaly. But in case the number is, for example, ten times or even hundred times higher, it might be a severe anomaly. This of course depends on how the grading rules are defined.

2.1.3 RULE MODELING BASED ANOMALY DETECTION

In a rule modeling based method rules are defined for the system and once these rules are broken, those instances are marked as anomalies. [10] Basically this is similar to how firewalls operate. Firewalls have predefined rules which are matched against network traffic. If the traffic is not in conflict with these rules it is then allowed to pass through. Everything that is against these rules is dropped. In anomaly detection this would mean that everything that is against the rules is thought of as an anomaly.

2.2 NETWORK INTRUSION DETECTION SYSTEM

From 21st century onwards, while networks have been developing rapidly, network based IDS has received more attention. Change of focus in IDS research from HIDS towards NIDS can also be explained by the research value. HIDS has been studied widely and new findings on that field are difficult to find. NIDS instead is a more interesting topic because network based intrusions are constantly increasing. This gives new opportunities for the research field to discover new methods that are able to detect previously unknown threats. Snort is an IDS/IPS that combines signature, protocol and anomaly based intrusion detection methods to efficiently detect and prevent intrusions. Snort has been developed by Sourcefire that also regularly provides rule updates to Snort. In addition to Snort, some network based IDS studies are discussed in the following paragraphs.

2.2.1 AUTONOMOUS AGENTS FOR INTRUSION DETECTION (AAFID)

AAFID was a project within the centre for education and research in information assurance and security (CERIAS) in Purdue University. The project group consisted of

students and faculty who were interested in developing a new type of intrusion detection system. Their approach is to use a distributed architecture of IDS agents to cover the operation of the whole network.

2.2.2 DISTRIBUTED SOFT COMPUTING INTRUSION DETECTION SYSTEM (D-SCIDS)

D-SCIDS consists of multiple distributed IDS sensors over a large network. IDSes communicate with each other directly or through a centralized server that also provides advanced network monitoring. In their research Abraham et al. [29] evaluated three fuzzy rule-based classifiers to detect intrusion in network and were then further compared with other machine learning techniques.

2.2.3 NEXT-GENERATION INTRUSION DETECTION EXPERT SYSTEM (NIDES)

NIDES are real-time IDS that monitor user activity on multiple target systems. NIDES are placed on a single host that analyses audit data collected from interconnected systems. Intrusion detection on NIDES is a hybrid of misuse detection and anomaly detection; a rule based signature analysis and a statistical profile-based anomaly detector. The notation expert in NIDES means a system that is intelligently processing intrusion alarms to decide whether further investigation from a security guard is needed or not. Further development of NIDES evolved into SRI's project called EMERALD.

2.3 CLASSIFIER CONSTRUCTION

Classifier construction is another important research challenge to build efficient IDS. Nowadays, many data mining algorithms have become very popular for classifying intrusion detection datasets such as decision tree, naïve Bayesian classifier, neural network, genetic algorithm, and support vector machine etc. However, the classification accuracy of most existing data mining algorithms needs to be improved, because it is very difficult to detect several new attacks, as the attackers are continuously changing their attack patterns. Anomaly network intrusion detection models are now using to detect new attacks but the false positives are usually very high. The performance of an intrusion detection model depends on its detection rates (DR) and false positives (FP). DR is defined as the number of intrusion instances detected by the system divided by the total number of the intrusion instances present in the dataset. FP is an alarm, which rises for something that is not really an attack. It is preferable for an intrusion detection model to maximize the DR and minimize the FP. For DR, we can modify the objective function to 1-DR. Therefore classifier construction for IDS is another technical challenge in the field of data mining.

III. LITERATURE SURVEY

A rich literature of intrusion detection focuses on feature reduction using soft computing and neural network. Many of these techniques are based on principal of component into a set of class classification problems. Despite the success of these techniques reported in different domains for various types of applications, such as text document classification, and speech recognition, most of these techniques are mainly proposed for learning from relatively balanced training data. However, in much application, the training data can be often intrusion, where some classes of data have a small number of samples compared to the other classes, and in which it is important to accurately classify the minority cases. In survey, numbers of anomaly detection systems are study based on many different

machine learning techniques. Some studies apply single agent learning technique, such as neural networks, genetic algorithms, support vector machines, etc. On the other hand, some systems are based on combining different learning techniques, such as hybrid or ensemble techniques. In particular, these techniques are developed as classifiers, which are used to classify or recognize whether the incoming Internet access is the normal access or an attack.

IV. PREVIOUS WORK DONE

We has studied various research and journal papers related to intrusion data classification. According to our research we have analyzed that many of the papers focuses on the problem of better classification of intrusion data and to use an optimized technique for it. Few review of summary described here and implicated with their respective author.

1. FAST FEATURE REDUCTION IN INTRUSION DETECTION DATASETS

In this paper author tried to propose a very simple and fast feature selection method to eliminate features with no helpful information on them. Result faster learning in process of redundant feature omission. We compared our proposed method with three most successful similarity based feature selection algorithm including Correlation Coefficient, Least Square Regression Error and Maximal Information Compression Index. After that we used recommended features by each of these algorithms in two popular classifiers including: Bayes and KNN classifier to measure the quality of the recommendations. There are varieties of attacks that IDS tries to detect. Some of these can be detected by scanning the packets to find signature of specific attacks. Other types of attacks are very much like normal packet pattern with slight difference in packet content. So there is two type of intrusion detection algorithm. They are called Misuse Detection and Anomaly Detection respectively.

2. INTRUSION DETECTION USING RANDOM FORESTS CLASSIFIER WITH SMOTE AND FEATURE REDUCTION

In this paper author described the feature reduction method with using classifier and the details are Synthetic Minority Oversampling Technique (SMOTE) is applied to the training dataset. A feature selection method based on Information Gain is presented and used to construct a reduced feature subset of NSL-KDD dataset. Random Forests are used as a classifier for the proposed intrusion detection framework. Empirical results show that Random Forests classifier with SMOTE and information gain based feature selection gives better performance in designing IDS that is efficient and effective for network intrusion detection. We used random forests classifier for developing efficient and effective IDS. For improving the detection rate of the minority classes (R2L and U2R) in imbalanced training dataset we used Synthetic Minority Oversampling Technique (SMOTE) and we picked up all of the important features of the minority class using the minority classes attack mode. Results from the experiment shows that our approach reduces the time required to build the model and also increases the detection rate for the minority classes in a considerable amount.

V. PROPOSED METHODOLOGY AND ARCHITECTURE

In this paper proposed a feature selection and reduction based intrusion detection system. The process of feature reduction and selection improved the detection and classification ratio of intrusion detection system. The feature selection process used

for find common feature for attacker participant and feature reduction process used for unwanted feature those who are not involved in attack and normal communication. for the reduction of feature used RBF neural network function. The RBF neural network function work on common feature correlation and generates similar and dissimilar pattern with the help of ACP algorithm. The reduction process reduces the large number of attribute and improved the detection of intrusion detection system. In the process of feature reduction various algorithm are used such algorithm are principle of component analysis and neural network. The reduction process used PCA method this method is static reduction technique, reduces only fixed number of attribute. The fixed number of feature reduction process not justify the value of feature it directly reduces the feature. On the consideration of computational time feature reduction is also an important aspects, the reduces feature increase the processing of detection ratio. Many methods have been proposed in the last decades on the designs of IDSs based on feature reduction technique. For example silakari and saliendra[4] proposed a generic framework for intrusion detection based on feature reduction and ensemble based classifier. On the other hand genetic algorithm is directly applied for classification in the work of Li [5]. Jain and Upendra [6] applied information gain based feature reduction for intrusion detection. They used KDDCUP'99 dataset for comparing four machine learning algorithms and they found that J48 classifier outperforms over BayesNet, OneR and NB classifiers. Muda et al. [7] also used KDDCUP'99 dataset for evaluating their K-Means and Naive Bayes based learning approach to carry out intrusion detection. Support Vector Machine (SVM) based IDS with Principal Component Analysis (PCA) dimension reduction is presented for intrusion detection in [8,9]. Z. Xue-qin et al. [10] proposed SVM IDS with Fisher score for feature selection. Zhang and M. Zulkernine [11] applied random forests for network intrusion detection. in this paper used ID3 algorithm for feature selection. ID3 is attribute based classification technique in decision tree. The selection of attribute in ID3 algorithm is entropy of information and gain of information. The increasing the sample selection area used radial biases function in ID3 algorithm. The continuity of chapter further discusses feature selection feature reduction, RBF function and finally discuss proposed algorithm.

5.1 FEATURES EXTRACTION

Intrusion detection systems can either have single variable approach or a multi-variable approach to detect intrusions depending on the algorithm used. In the single variable approach a single variable of the system is analyzed. This can be, for example, port number, CPU usage of a local machine etc. In multi-variable approach a combination of several features and their inter-correlations are analyzed. [] In addition based on the method the way in which features are chosen for the IDS can be divided into two groups; into feature selection and feature reduction.

VI. IMPLEMENTATION

In this paper we perform experimental process of proposed Feature reduction algorithm for intrusion detection system. The proposed method implements in mat lab 7.8.0 and tested with very reputed data set from UCI machine learning research center. In the research work, I have measured detection accuracy, true positive rate, false positive rate, true negative rate and finally false negative rate error of classification method. To evaluate these performance

parameters I have used KDDCUP99 datasets from UCI machine learning repository [42] namely intrusion detection dataset.

CONCLUSION

In this thesis proposed a feature based intrusion data classification technique. This reduces feature improved the classification of intrusion data. The reduction process of feature attribute performs by RBF function along with feature correlation factor. The proposed method work as feature reducers and classification technique, from the reduction of feature attribute also decrease the execution time of classification. The decrease time increase the performance of intrusion detection system. Our experimental process gets some standard attribute set of intrusion file such as pot_type, service, sa_srv_rate, dst_host_count, dst_host_sa_srv_rate. These feature attribute are most important attribute in domain of traffic area. The classification rate in these attribute achieved 98 %.

REFERENCES

[1] Shafigh Parsazad, Ehsan Saboori, Amin Allahyar "Fast Feature Reduction in Intrusion Detection Datasets" MIPRO 2012, Pp 1023-1029.

[2] Abebe Tesfahun, D. Lalitha Bhaskari "Intrusion Detection using Random Forests Classifier with SMOTE and Feature Reduction" International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, 2013. Pp 127-132.

[3] Hachmi Fatma, Limam Mohamed "A two-stage technique to improve intrusion detection systems based on data mining algorithms" IEEE, 2013. Pp 1-6.

[4] Shailendra Singh, Sanjay Silakari "An Ensemble Approach for Cyber Attack Detection System: A Generic Framework" 14th ACIS, IEEE 2013.

[5] Li, "Using Genetic Algorithm for Network Intrusion Detection" Proc. the United States Department of Energy Cyber Security Group 2004 Training Conference, May 2004.

[6] Jain , Upendra "An Efficient intrusion detection based on Decision Tree Classifier using feature Reduction", International Journal of scientific and research Publications , Vol. 2, Jan. 2012.

[7] Dewan Md. Farid, Jerome Darmont, Nouria Harbi, Nguyen Huu Hoa, Mohammad Zahidur Rahman "Adaptive Network Intrusion Detection Learning: Attribute Selection and Classification" 2008. Pp 1-5.

[8] Gary Stein, Bing Chen, Annie S. Wu, Kien A. Hua "Decision Tree Classifier For Network Intrusion Detection With GA-based Feature Selection" 2556. Pp 1-6.

[9] Ritu Ranjani Singh a, Prof. Neetesh Gupta "To Reduce the False Alarm in Intrusion Detection System using self Organizing Map" in International journal of Computer Science and its Applications.

[10]Z. Xue-qin, G. Chun-hua, L. Jia-jin "Intrusion detection system based on feature selection and support vector machine" Proc. First International Conference on Communications and Networking in China (ChinaCom '06), Oct. 2006.

[11] Zhang , M. Zulkernine "Network Intrusion Detection using Random Forests" School of Computing Queen's University, Kingston Ontario, 2006.

[12] John Zhong Lei and Ali Ghorbani "Network Intrusion Detection Using an Improved Competitive Learning Neural Network" in Proceedings of the Second Annual Conference on Communication Networks and Services Research IEEE.

[13] P. Jongsuebsuk,N. Wattanapongsakorn and C. Charnsripinyo "Network Intrusion Detection with Fuzzy Genetic Algorithm for Unknown Attacks" in IEEE 2013.