Analysis of Feature Selection and Ensemble Classifier Methods for Intrusion Detection

H.P. Vinutha, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India
Poornima Basavaraju, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India

ABSTRACT

Day by day network security is becoming more challenging task. Intrusion detection systems (IDSs) are one of the methods used to monitor the network activities. Data mining algorithms play a major role in the field of IDS. NSL-KDD’99 dataset is used to study the network traffic pattern which helps us to identify possible attacks takes place on the network. The dataset contains 41 attributes and one class attribute categorized as normal, DoS, Probe, R2L and U2R. In proposed methodology, it is necessary to reduce the false positive rate and improve the detection rate by reducing the dimensionality of the dataset. Use of all 41 attributes in detection technology is not good practices. Four different feature selection methods like Chi-Square, SU, Gain Ratio and Information Gain feature are used to evaluate the attributes and unimportant features are removed to reduce the dimension of the data. Ensemble classification techniques like Boosting, Bagging, Stacking and Voting are used to observe the detection rate separately with three base algorithms called Decision stump, J48 and Random forest.

KEYWORDS

Accuracy, Bagging, Boosting, Dimensionality Reduction, DoS, Ensemble Techniques, Feature Selection, Intrusion Detection System, NSL-KDD Dataset, Probe, R2L, Stacking, U2R, Voting Detection Rate

INTRODUCTION

The antivirus software, message encryption, password protection, firewalls, secured network protocols, etc., are not enough to provide security in network. Intruders may attack the network using many unpredictable methods. Monitoring the activities and actions on network systems or a computer network using Intrusion Detection Systems (IDSs) helps to analyze the possible incidents. In 1980 the first IDS was launched by James P Anderson and later in 1987 D Denning enhanced it. IDSs play a major role in data integrity, confidentiality and availability of the network. Some of the attacks included in the intrusion are confidential and sensitive information malign, available resources and functionalities may be hacked. Means and modality logs of various attacks are easily produced by IDS. This helps to prevent possible attacks in future. Current day’s organizations are provided with good source of security analysis by using IDS. IDS divided into two categories based on their architecture.

DOI: 10.4018/IJNCR.2018010104
and functionality. The categories of IDS are host-based intrusion detection system (HIDS) and network-based intrusion detection system (NIDS). HIDS are running on individual host machine and NIDS are within the network to monitor to and from traffic on the network. This proposed work we have concentrated on NIDS in order to monitor the network traffic to identify the incoming traffic is normal or anomaly. We have two important modes and methods to operate IDS for packet analysis on the network they are anomaly detection and misuse detection methods.

**Anomaly Detection**

Anomaly detection is a method of scanning for the abnormal activities that encounter on the network. It maintains the log of activities that takes place on the network and such information can be used for the comparison of all the activity which takes place on the network. Using these information new rules can be defined for the kind of new activity, if any deviations takes place from the normal activity can be referred as an anomaly. Some of the common examples for rule-based methods are Minnesota Intrusion Detection System (MINDS), Intrusion Detection Expert System(IDES), Next Generation Intrusion Detection Expert System (NGIDES), etc.

**Misuse Detection**

Another method of IDS is misuse detection. In this method the network activities are compared with pre-defined signatures. Signatures are some set of characteristic features that gives specific patterns of attack. These set of characteristics features are stored to compare with the network activity, if any pattern is different from the pre-defined patterns can be considered as attacks. Some of the common signature-based tools used are Snort, Suricata, KISMAT, HONEYD, BRO Ids etc.

Current IDSs are encountered with many drawbacks some of the major drawbacks are identified as false positive and false negatives. False positive occurs when normal is considered as malicious attack. In false negative encounters when actual attack take place. Data mining is one of the field contributed many techniques for IDSs. Techniques like data summarization, visualization, clustering, classification etc helps to accomplish the task. Along with these drawbacks IDSs is facing major drawback in Big Data because in this huge volume of data has to be managed.

**Detection Issues**

There are four main type of detection issues in IDS and these issues rises depending on the type of alarm in the intrusion scenario. IDSs encounter the falling type of detection issues:

- **True Positive**: IDS response to raise alarm when actual attack occurs.
- **True Negative**: IDS does not response to raise an alarm when no attack occurs.
- **False Positive**: IDS generate an alarm when no attack takes place.
- **False Negative**: IDS does not generate an alarm when actual attack takes place.

Intrusion detection system is becoming very challenging task in current days. The dataset used for IDS will be huge in number and contains many irrelevant features and sometimes redundant relevant features. In stage of detection of intrusion if our system makes use of all the features in the dataset, analysis of intrusion becomes very difficult. Because if dataset contains large number of features then this makes it difficult to identify the suspicious behaviors. In such cases it is going to reduce the detection performance and efficiency of the model. So, it is necessary to reduce the dimension of the dataset before applying the data mining approaches such as classification, clustering, association rule and regression on the dataset. Feature selection methods are used as pre-processing step to reduce the dimensionality of dataset.
Evolutionary Modeling and Industrial Structure Emergence
[www.igi-global.com/chapter/evolutionary-modeling-industrial-structure-emergence/21135?camid=4v1a](www.igi-global.com/chapter/evolutionary-modeling-industrial-structure-emergence/21135?camid=4v1a)

Quorum Sensing Digital Simulations for the Emergence of Scalable and Cooperative Artificial Networks