A Review on Intrusion Detection System for HTTP Based Services

Susan Rose Johnson, Anurag Jain

Abstract— Nowadays, internet has become the main source for communication. Various organizations, companies, universities, schools provide their services through network facility. As a result, security has become a major issue in the field of networking. An Intrusion Detection System (IDS) is used to monitor the intrusions over the network and classify them as normal or attack class. In this paper, IDS is used along with the Support Vector Machine (SVM) and random forest for the classification of network traffic as normal or attack. The main aim is to enhance the performance of IDS by preparing the training dataset that detects the malicious attacks which misuse the http services. NSL – KDD dataset has been used to train and test the IDS. A feature selection technique called random projection is used to select the relevant features from the dataset. SVM and random forest jointly classifies the unknown data. Finally, the result is optimized using Ant Colony Optimization (ACO). The main goal of this algorithm is to find an optimal path in the graph based on the behavior of the network.

Index Terms— Network Security;Intrusion Detection System (IDS); Support Vector Machine (SVM), Ant Colony Optimization (ACO).

1 Introduction

THESE days, network is considered as the vital part of communication. Most of the services are processed through computer network over internet. The tremendous growth in network and accessibility of the internet has gained a lot of attention amidst individuals. The information that we want to send should be secured. The security of confidential information is one of the major challenges to be faced in networking world. As the numbers of network attackers are increasing gradually, it is important to secure the confidential information over the internet. For this purpose various security tools have been used nowadays. Some of them are firewalls, anti – virus software's, Intrusion Detection System (IDS) and so on. Firewalls act as a barrier between the internal network and the outside world. It can only detect the boundary level attacks. So, it is not worth enough to detect internal attacks. IDS is a device or software application mainly used to observe the intruder events over the network. After monitoring they classify the network traffic into attack class or normal class. It acts as an alert system that reports when an unauthorized activity is detected by the system. The accuracy of the IDS depends upon detection rate. If high performance of the IDS, then it means that the accuracy of detection is high. In this work, SVM and random forest based IDS has been proposed to detect the intrusion over the network. A feature selection technique has been used to select the relevant features from a set of data items. The data items consist of various features or attributes that represent the network connection. So, the irrelevant and redundant features can be removed from the feature selection process. Thus, it minimizes the processing delays.

- *Author name is currently pursuing masters degree program in electric power engineering in University, Country, PH-01123456789. E-mail: author_name@mail.com*
- *Co-Author name is currently pursuing masters degree program in electric power engineering in University, Country, PH-01123456789. E-mail: author_name@mail.com*
(*This information is optional; change it according to your need.*)

2 A Brief Review on IDS

An intelligent Intrusion Detection System (IDS) can be built using data set. A data set is a collection of data that is used to evaluate the performance of IDS. In this work, NSL - KDD data set is used by the IDS. The NSL - KDD data set has been used over years to demonstrate the applicability and performance of the knowledge discovery techniques. They are used to analyze the performance of the classification algorithm used for detecting the anomalies in the network connection of IDS. The NSL- KDD data set consists of 41 features and 5 classes in which 4 classes represent the attack types and the other is the normal class.

The attack types are:

- DoS.
- Probe.
- R2L.
- U2R.

The benefits of NSL-KDD data set are:

- No redundant records are included in the training set.
- No duplicate records in the test set.

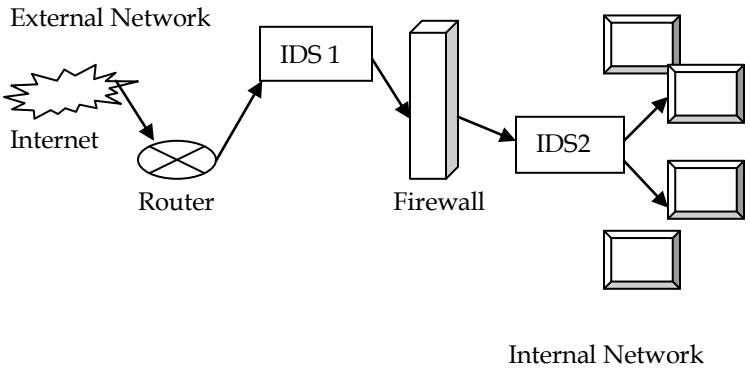

Fig1. Intrusion Detection System

A feature selection technique has been used to select the relevant features from a set of data items. The data items consist of various features or attributes that represent the network connection. So, the irrelevant and redundant features can be removed from the feature selection process. Thus, it minimizes the processing delays. This paper uses random projection as the feature selection technique. The random projection technique is used to reduce the dimensionality of a set of points. As the name suggests, it reduces the number of random variables. Thus the problem of managing the large datasets can be reduced. In random projection, the original d-dimensional data is projected into k-dimensional (k<<d) subspace through origin. Classification is a process used for discovering classes of unknown data. Some of the classification techniques are:

- Rule based classifier.
- Naïve Bayes classifier.
- Support Vector Machine.
- ANN.
- Random Forest.
- Decision Tree and so on.

The list of features of NSL – KDD data set is given below in the table:

Table 1 List of Features

Feature No.	Feature
1	duration
2	protocol_Type
3	service
4	flag
5	src-bytes
6	dst_bytes
7	land
8	wrong_fragment
9	urgent
10	hot
11	num_failed_logins
12	logged_in
13	num_compromised
14	root_shell
15	su_attempted
16	num_root
17	num_file_creations
18	num_shells
19	num_access_files
20	num_outbound_cmds
21	is_hot_login
22	is_guest_login
23	count
24	srv_count
25	serror_rate
26	srv_serror_rate
27	rerror_rate
28	srv_error_rate
29	same_srv_rate
30	diff_srv_rate
31	srv_diff_host_rate
32	dst_host_count
33	dst_host_srv_count
34	dst_host_same_srv_rate
35	dst_host_diff_srv_rate
36	dst_host_same_src_port_rate
37	dst_host_srv_diff_host_rate
38	dst_host_serror_rate
39	dst_host_srv_serror_rate
40	dst_host_rerror_rate
41	dst_host_srv_error_rate

The classification process mainly consists of two phases, training and testing phase. During the training phase, a classification model is built from train set. During the testing phase, the model is evaluated using test set.The selection of a classification technique is a challenge in the field of network security, because the selected technique should provide accurate result with less time consuming. In this paper, Support Vector Machine (SVM) and random forest method is used to classify the upcoming data. Support Vector Machine (SVM) or Support Vector Network is a supervised learning technique that analyzes data used for classification. There will be predefined data for the classification of unknown data. The predefined data are called as training samples which are used to train the IDS for the classification process. After analyzing, the unknown data would be placed in either of two classes. An optimal separating hyperplane is used to accurately classify the unknown data. The optimal hyperplane is obtained by support vectors i.e. training records and margins.

The advantages of Support Vector Machines are:

- SVMs are effective in high dimensional spaces.
- SVMs work well even if the training samples are very less.
- SVMs are flexible and powerful.
- SVMs produce accurate results.

The disadvantages of Support Vector Machines are:

- If the number of features is more than samples, poor performance would be the result.
- There is no multi-class SVM, we must combine two SVMs.
- More time consuming during training phase.

The random forest method is an ensemble approach that is based on the principle divide – and – conquers method which is used in the classification task. As it is an ensemble method, it combines a group of weak learner to produce strong leaner which can classify the data accurately. They combine the bagging idea and random selection of features. Each tree in the random forest represents classes such as normal and different attack classes.

The advantages of Random Forest are:

- Random forest can run efficiently on large databases.
- Random forest handles an N number of input variables without variable deletion.
- It provides the most important features in the classification.

- It can perform well even if the data are missing.

The limitations of Random Forest are:

- The large number of trees in random forest can cause delay in processing.
- It has been noted that random forest suitable for some datasets.

The Ant Colony Optimization (ACO)/ swarm algorithm is used for the optimization process. It is a probabilistic technique used to solve the computation problems. The main goal of this algorithm is to find an optimal path in the graph based on the behavior of the network.

3 Literature Survey

This section gives an extensive literature survey on the various IDS used to detect the malicious connection. We study on several research papers and journals and gain knowledge on IDS. All methodology and process are not described here. But some related works in the field of IDS are discussed.

Mohamed M. implemented IDS based naïve bayes classifier [1] to sort the connection as normal or attack. The first proposed IDS was implemented on NSL – KDD dataset where the detection rate seems to be low. As a result, the IDS is applied to http traffic and classifies the http service attacks as: Neptune, Back, Portsweep, Saint and appache2. Thereafter, it shows a good performance, but it may increase the computational time.

Hari O et al. proposed the anomaly detection IDS [2]. It contains various classification algorithms such as k-means algorithm, k-nearest neighbour classifier and naïve bayes classifier. Entropy based algorithm is used to select the important features that can accurately identify the attack or intrusions over the network. After the feature selection process, the k – means algorithm is used for clustering. This system can achieve 98.18% detection rate and the false positive rate is 0.83%. The correctness of classifying the attacks range from 92% to 98%.

The positive aspect is that they can detect the attacks and categorizes them as :

- DoS.
- Probe.
- U2R.
- R2L.

The negative aspect is that they are time consuming. Because the concepts such as k-means, k-nearest neighbour classifier and naïve bayes classifier were used as the classification process.

Chirag M et al. implemented IDS based on Snort and Bayes theorem in the cloud environment [3]. Snort is a network intrusion detection and prevention system. It uses a rule-based language that joins signature, protocol and anomaly inspection techniques. The packet sniffer in snort captures the packets from the network using the signature based detection technique. The captured malicious packets are stored in the log database and the remaining packets are checked using the Bayesian classifier. The behaviors of the packets are analyzed by the naïve bayes classifier and finally they classify the intrusion packets and normal packets. Therefore, the normal packets are permitted into its destination host system and the intrusion packets are logged into the database. The detection rate of the IDS is 96% and the false positive rate is 1.5%. The negative aspect of this work is that since they are using two filtering techniques, it may lead to processing overhead causing delay.

Chandrasekhar M et al. implemented Data Mining techniques [4] for the classification of intrusions or malicious connections over the internet. The techniques are k-means clustering, Neuro-fuzzy training, SVM and Radial SVM. The k - means clustering is used to cluster the incoming data set into five clusters. Out of five clusters, one cluster represents the normal behavior and the remaining four clusters represent the intrusion behavior. The Neuro fuzzy classifier is associated with each cluster. They are trained using data contained in each cluster. So, the normal cluster contains the data that can describe the normal behavior only. Likewise, the attack cluster contains only the data relevant to the attack behavior. SVM is used to decrease the attributes. As we all know that feature selection is a significant task in IDS. It is necessary to select the features accurately from a list of data sets. This task simplifies the classification process and improves the efficiency of the system. Finally, the radial SVM detects the intrusions over the internet. The detection rate of the system is 97.5%. Since, it uses above mentioned data mining techniques this is a time consuming process.

The negative aspect is that they are time consuming. Because the concepts such as k-means, k-nearest neighbour classifier and naïve bayes classifier is used are used as the classification process.

T. Alexander et al. proposed Shiryaev–Roberts procedure [5] to observe the attacks as soon as it is happening. They reduce the detection delays through this procedure. Comparing with other similar procedures, this procedure is inexpensive and can be easily implemented in the real time IDS. The negative aspect of this procedure is it increases the false alarm rate. So, it requires extra filtering technique with high detection accuracy which in turn increases the processing delay.

Sumaiya T et al. introduced Tree based IDS classification algorithms [6]. It shows that the detection rate is 97.49% with a false detection rate is 2.5%. Here, it clearly shows the detection rate needed to be enhanced. At the same time, the false detection rate also must be enhanced. This shows that the work requires enhancement.

Dhanya Jayan et al. introduced detection of malicious client based HTTP/DoS attack on web server [7]. A web server side defense system has been implemented against the HTTP attack. The aim of the proposed system is to identify the malicious connections that attack the HTTP services. The request from the client to the web server has been observed. The proxy server acts as an intermediate between the client and the server. The request can be of two types: attack or normal. There

are two phases mainly: attack learning phase that maintain attack request based on the certain parameter values and the detection phase detects the http/DoS attack by comparing the parameter values and term vector. Finally, the attack requests are grouped into separate_attack_sequence.

Y. Xie et al. introduced web proxy's behavior based on locality principles for http attack [8]. In this paper, the malicious web requests are detected by using the locality behavior of web proxy. For the detection of http attacks, the behavior of web proxy has been observed. The behavior includes spatial and locality behavior access for the accurate detection of http attack. The disadvantage is that it doesn't include the spatial behavior for the attack detection.

P. Garcia et al. proposed anomaly based network intrusion detection [9]. The system uses hidden markov model for the detection of the http attack in the network. The hidden markov model consists of observable states as well as the hidden states. As some of the states are hidden we cannot identify the attack due to lack of information and the computational complexity of obtaining the parameters using markov model is considered as the drawbacks.

Yi. Xie et al. introduced web proxy based http attacks by using temporal and spatial behavior [10]. In this paper, the current behavior is compared with the predefined behavior for the http attack. The Gaussian – Gamma – Hidden – semi - Markov Model (GGHsMM) has been implemented to detect the http attack. They make use of TSL to analyze the proxy behavior. Since, the system uses GGHsMM the information regarding attack couldn't be fulfilled. The detection rate is low.

Sudhansu Ranjan introduced HTTP service based NIDS in cloud computing [11]. In this paper, naïve bayes classifier has been implemented to detect the intrusions over the network.
The positioning of the IDS is an important aspect to be considered while talking about the IDS. So, they explained the positioning of IDS. There are mainly three phases: training, testing and execution. During the training phase, the naïve bayes classifier classifies the data set into normal class or attack class. In testing and execution phase, the unknown network traffic is classified based on the probability values. The snort algorithm is used to check the captured packets with the predefined rules. The detection rate is 99.75% and false positive rate is 0.54%.

4 Proposed Methodology

The proposed methodology will use NSL-KDD data set to measure the performance of IDS. It is an improved version of KDDCup'99 data set. Due to inefficiency in KDD data set discovered by many researchers, we go for NSL-KDD data set. It contains 41 features of a network connection that include training and testing data. It is important to select the relevant features from the data set. Here, an efficient algorithm is used for feature selection, i.e. random projection technique. The feature selection techniques are used to select a subset of relevant features for the model construction. The main benefits of using a feature selection technique include the simplification of models that can be easily analyzed by the users or researches; they reduce the number of features and select only the relevant attributes for the classification process which in turn decreases the processing time.

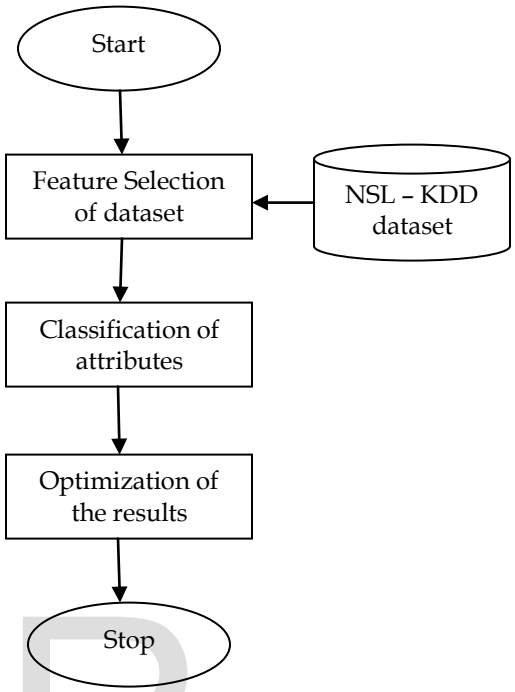

Fig.2 Proposed Method

After that we use two classification technique SVM and random forest. The classification process is used to assign the unknown traffic into a particular class based on some predefined train data. The algorithm tries to obtain the relationship between the attributes so as to classify the unknown network into either normal class or attack class. The random forest technique combines multiple decision trees to obtain a good classification result. Support Vector Machine (SVM) is a supervised learning technique that analyzes data used for classification. With the help of training samples it sort the upcoming data into various classes. Then we try to optimize the results using optimization techniques such as ACO or swarm optimization. The main goal of this algorithm is to find an optimal path in the graph based on the behavior of the network.

5 Comparison of Various IDS

In this section comparison among different IDS has been done. Comparisons of IDS's are mainly done to evaluate the performance of the IDSs while they are subjected to various techniques. Finally, we will be getting the comparison results. The efficiency of IDS depends upon many factors such as feature selection technique that can offer better features for the accurate identification of miscellaneous data as well as the normal data; the algorithms selected for classification / clustering of the upcoming data. These are the factors that are to be made significant while building IDS. Given below is a comparison table that shows the performance of IDSs,

Table 2 Comparison of Various IDS

IDS Index	Intrusion Detection Systems	Detection Rate (%)
I	IDS based on Snort and Bays theorem [3]	96
II	Tree based IDS [6]	97.49
III	IDS based on DM [4]	97.5
IV	k-means, k-nearest neighbor and naïve bayes classifier [2]	98.18
V	HTTP service based IDS [1]	99.03

6 Conclusion

In this work, IDS based on SVM and random forest has been implemented to classify the data as normal or attack. To evaluate the performance of IDS, the NSL – KDD dataset has been used. The IDS is trained and tested using dataset. The proposed work implemented random projection as the feature selection technique. It is considered as one of the best and efficient feature selection technique amidst other feature selection technique. They combine the features to produce a good output. The classification technique SVM and random forest simultaneously classified the unknown data with a high detection rate. The Ant Colony Optimization/ swarm optimization maximized the detection rate.

Acknowledgment

This work is partially supported by Prof. Anurag Jain Supervisor, and Head (Department of Computer Science). His valuable advice and support, was an inspiring force to me. I would like to thank Dr. Yuvraj Rana, Director and Dr. J. L. Rana, Group Director for his valuable guidance and motivation.

References

[1] Mohamed M. Abd-Eldayem "A proposed HTTP service based IDS" in Egyptian Informatics Journal, Volume 15, Issue 1, March 2014.

[2] Hari O, Aritra K. "A hybrid system for reducing the false alarm rate of anomaly intrusion detection system". In: 1st IEEE international conference on recent advances in information technology, RAIT 2012. Dhanbad, India; March 2012.

[3] Chirag M, Dhiren P. Bayesian classifier and snort based network intrusion detection system in cloud computing. In: The third IEEE international conference on computing communication & networking technologies, ICCCNT 2012. Coimbatore, India; July 2012.

[4] Chandrasekhar M, Raghuveer K. Intrusion detection technique by using k-means, fuzzy neural network and SVM classifiers. In: IEEE international conference on computer communication and informatics. Coimbatore, India; January 2013.

[5] T. Alexander, P. Aleksey, Grigory So "Efficient computer network anomaly detection by change point detection methods". IEEE J Sel Top Signal Process, 7 (1) (2013).

[6] Sumaiya T, Aswarmi C. "An analysis of supervised tree based classifiers for intrusion detection system". In: IEEE proceedings of the international conference on pattern recognition, informatics and mobile engineering, PRIME 2013. Salem, India; February 2013.

[7] Dhanya Jayan, Pretty Babu "Detection of Malicious Client based HTTP/DoS" in IJSR volume 3 Issue 7, July 2014.

[8] Y. Xie and S. Yu, Measuring the Normality of Web Proxies Behavior Based on Locality Principles, Network and Parallel Computing, vol. 5245, pp. 61-73, 2008.

[9] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Macia-Fernandez, and E. Vazquez, Anomaly-Based Network Intrusion Detection: Techniques, Systems and Challenges, Computers and Security, vol. 28, nos. 1/2, pp. 18-28, 2009

[10] Yi Xie, S. Tang, Y. Xiang and J. Hu, Resisting Web Proxy-Based HTTP Attacks by Temporal and Spatial Locality Behavior, IEEE transactions on parallel and distributed systems, VOL. 24, NO. 7, JULY 2013.

[11] Sudhansu Ranjan "HTTP Service based Network Intrusion Detection System in Cloud Computing" IJSRMS, volume 1, Issue 1, August 2015.R.J. Vidmar, "On the Use of Atmospheric Plasmas as Electromagnetic Reflectors," *IEEE Trans. Plasma Science*, vol. 21, no. 3, pp. 876-880, available at http://www.halcyon.com/pub/journals/21ps03-vidmar, Aug. 1992. (URL for Transaction, journal, or magzine)