An Improved Intrusion Detection System using Random Forest and Random Projection

Susan Rose Johnson, Anurag Jain

Abstract— Communication plays a significant role in everybody's life. Computer network is a type of communication network where information can be passed from one individual to another. Most of the services of various organizations, companies, universities, schools are processed via internet. Computer network was developed with a motive to make communication easier amidst individuals. The remarkable growth in network and ease of access to internet increased the security issues in the field of network. As the number of network attacks has increased, the security issues have been affected over past few years. An Intrusion Detection System (IDS) is used to watch unauthorized activities through the network. It can organize the unidentified records as regular class or attack class. In this paper, two classifiers random forest and support vector machine along with random projection is implemented. Then the same classifiers are used without the help of random projection. These classifiers classify the unidentified data into attack classes such as probe, DoS, U2R, R2L. The NSL – KDD dataset is used to assess the IDS. The random projection procedure is used to choose the applicable attributes from the data set after which the classifiers random forest and SVM classifies the unidentified data. The performance of IDS is evaluated by the data set. A comparison between the classification algorithms has been made. The outcome shows that the detection rate of random projection along with random forest is approximately 100% which is more efficient than random projection and SVM, random forest and SVM without random projection.

Index Terms— Network Security; Intrusion Detection System (IDS); DoS; U2R; R2L; KDD; Support Vector Machine (SVM)

1 INTRODUCTION

FOR the past few years, network plays a significant role in communication. The computer network allows the computing network devices to exchange information from individuals to individuals. The services of various organizations, companies, colleges, universities are accessed throughout computer network. This leads to a massive growth in networking field. The accessibility of internet has acquired a lot of interest among individuals. In this context, security of information has become a great challenge in this modern era. The information or data that we would like to send be supposed to be secured in such a way that a third party should not take control over them. When we are talking about security, we have to keep three basic factors in our mind: Confidentiality, Integrity and availability. Confidentiality means privacy of information. It gives the formal users the right to access the system via internet. This can be performed suitably along with accountability services in order to identify the authorized individuals. The second key factor is integrity. The integrity service means exactness of information. It allows the users to have self-assurance that the information passed is acceptable and has not been changed by an illegal individual. Availability service means information to be useful. An individual can access computer system; the information contained on the system. The availability also allows the computer system to pass on the information from one place to another. As the demand for internet increased the number of network attackers also increased. Now days, quite a lot of defense tools have been used for security purpose such as firewalls, anti – virus software's, Intrusion Detection System (IDS) and so on. Firewall is a network security system that observes and controls the unauthorized activity over the network. It acts as a hurdle between the in-house network and outdoor network. But, the drawback of using firewall is that it can only detect border line attacks. So, it is not useful enough to notice the internal attacks. The security of secret information is a significant challenge faced now days. We all know that the statistics of network intruders are rising each day, so it is essential to protect the private information that has to be sending over internet. It allows the unauthorized users to have entry on information over internet. For this purpose Intrusion Detection System (IDS) is a good option. IDSs are tools mainly used to examine the intruder events taking place via network. It acts as an alarm system that is capable of reporting an illegal activity when detected. After examination they order the network interchange into attack class or normal class. The exactness of the IDS depends upon finding rate. If the performance is high for the IDS, then the correctness of detection is also high. In this paper, SVM and random forest along with random projection is implemented. IDS have been anticipated to identify the intrusions over the network. A feature selection procedure has been implemented to choose the applicable features. The records consist of a range of attributes that characterize the network connection. So, the inappropriate and unnecessary attributes can be deleted by the pre - processing. As an end result, the processing time would be decreased.

Rest of the paper is organized as follows: in section 2 a description on intrusion detection system is presented with all necessary details, in section 3 the brief description on random projection followed by support vector machine in section 4 which is then followed by random forest in section 5. In section 6 a description on the proposed work is made. In section 6 experimental results followed by conclusion.

International Journal of Scientific & Engineering Research Volume 7, Issue 10, October-2016
ISSN 2229-5518

2 IDS TECHNOLOGY

An Intrusion Detection System (IDS) is used to watch malicious activities over the network. It can sort the unfamiliar records as normal or attack class. First monitoring is done, and then they order the network traffic into malicious class or regular class. It acts as an alarm system that reports when an illegal activity is detected. The exactness of the IDS depends upon detection rate. If the performance is high for the IDS, then the correctness of detection is also high. In truth, some of the intrusion detection systems are marketed with the ability to stop attacks before they are successful. We are even seeing new systems marketed as intrusion detection systems. Intrusion detection systems have existed for a long time. They are used to shield an association from attack. It is a relative concept that tries to identify a hacker when penetration is attempted. Ideally, such a system will only alarm when a successful attack is made. Intrusion detection system is not a perfect solution to all attack types. A good security program or security tools cannot be replaced by an IDS. The genuine users who may try to access the information for their pleasure cannot be identified with the help of IDS. The goals of IDS provide the requirements for the IDS policy. The potential goals include the following:

- IDS detect attacks.
- IDS traces user activity from point of entry to
- IDS generate alerts when required.
- Detect errors in system configuration.
- Provides security of the system without the need of non – expert staff.
- IDS can detect when the system is under attack.
- Provides evidences for attack.

A quick Intrusion Detection System (IDS) can be built using data set. A data set is a collection of data that is used to measure the performance of IDS. In this paper, NSL - KDD data set is used by the IDS. The NSL - KDD data set has been used over years to show the efficiency and concert of the knowledge discovery techniques. It is used to analyze the performance of various algorithms used for detecting the malicious activities over network. The NSL- KDD data set consists of 42 features where the 42nd feature is the attack label. The attack label consists of four types of attack. They are:

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

The benefits of NSL-KDD data set are:

- Unnecessary records are not included in the training set.

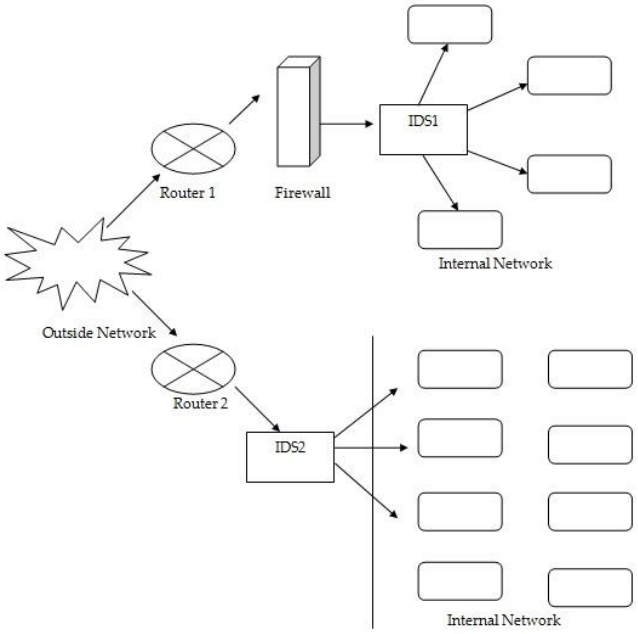

Fig1. Intrusion Detection System

- Duplicate records are not available in the test set.

IDS can be grouped into two: Network Intrusion Detection System (NIDS) and Host Intrusion Detection System (HIDS). Network Intrusion Detection system is a mechanism that is used within the network to identify the malicious event. The network traffic is monitored in the network that is in the subnet. If an attack is observed it matches the traffic with the known attack list. Then an alert is passed to the administrator. NIDS is installed in router to identify the passage of network traffic. HIDS runs on an individual system. The functions of two IDSs are the same. HIDS also monitors the unauthorized activity. It takes a short review of the existing files in the system. Then it matches it with the old system files. If it finds an intrusion or changes in the system, then an alert is passed to the administrator. The intrusion can be detected as if a file is modified or deleted, then it means malicious activity is reported.

The advantages of Network Intrusion Detection System (NIDS) are:

- Large networks are observed by installing fewer devices.
- NIDS can be fully hidden on the network.
- NIDS can arrest the inside contents of all packets travelling from one system to other system.

Disadvantages of NIDS are:

- NIDS can only alarm if the traffic matches the predefined rules.
- NIDS cannot determine if the attack was success-

- *Susan Rose Johnson is currently pursuing masters degree program in computer science and engineering in RITS Bhopal, India, PH-07691948398. E-mail:susanrose17@gmail.com*
- *Anurag Jain is Head of Department in computer science and engineering RITS Bhopal, India, PH-09425600625. E-mail: anurag.akjain@gmail.com (This information is optional; change it according to your need.)*

International Journal of Scientific & Engineering Research Volume 7, Issue 10, October-2016
ISSN 2229-5518

ful.
- NIDS cannot examine the traffic that is encrypted.
- Switched networks require special configurations so that the NIDS can monitor all the traffic.

3 Random Projection

The random projection technique is used to ease the dimensionality of a group of points. From the name itself, it reveals that it reduces the amount of random variables. Thus the difficulty of organization of large datasets can be reduced. In random projection technique, the original d-dimensional data is projected to k-dimensional (k<<d) subspace all the way through origin. The group of points lies in the Euclidean space. Random projection methods are dominant methods known for their effortlessness and less incorrect results compared with other methods. According to investigational outcome, random projection conserve sound distances, but experimental outcomes are sparse. This method is effortless and computationally well-organized task made to reduce the dimensionality of records by trading a controlled quantity of fault for quicker processing times and lesser model sizes. The dimensions and distribution of matrices are controlled in order to preserve the pair wise distances between any two records.

4 Support Vector Machine

Support vector machine is a technique that has been emerged for the analysis of data for the classification process. The support vector machine is also known as support vector networks. The SVM uses a set of training data where each one has been labeled into one of two categories. The training data set builds a model and the new unknown data would be categorized into the proper group. There will be a linear separation between the data that has to be classified. With the help of this line the data can be easily separated with more accuracy. In this technique, two categories are available; we can either classify the data to one class or to the other class depending upon the behavior of the new data. In addition to linear classification, they can perform non linear classification also. In non linear classification, data are not labeled so supervised learning is not possible. In such a context, unsupervised learning approach is implemented, which attempts to cluster the data with similar behavior.

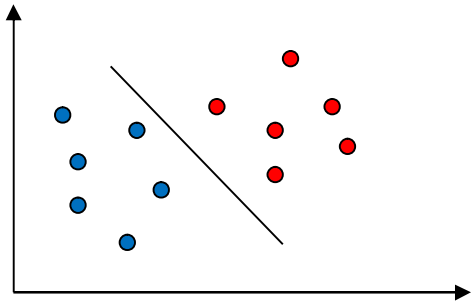

Fig2. Classification of data by SVM

Support Vector Machines (SVM) can be used for both classification as well as regression. However, it is mostly used for the classification problems. We perform classification based on the hyper plane i.e. the separation line that differentiates the two categories. The most important thing while using SVM is that how we can obtain the correct hyper plane. The accurate hyper plane can be obtained by using thumb rule.
The benefits of Support Vector Machine (SVM) are:
- SVM can generate correct and strong classification results.
- SVM are flexible and powerful classification algorithm.
- It can produce best results even if the training data are few in number.

The drawbacks of SVMs are:
- Speed and size of training and testing is the main issue in SVM.
- Lack of transparency of the results.
- We must combine two SVMs as there is no multi-class SVM.

5 Random Forest

Random forest method is ensemble methods which is based on the rule divide – and – conquer scheme used in the classification mission. As it is an ensemble process, it amalgamates a group of fragile learner to produce well-built leaner which can categorize the data precisely. They unite the bagging scheme and random selection of features. Random forest produces n number or tress. Each tree in the random forest represents regular and different malicious classes. It can run on large number of data sets.

The benefits of Random Forest are:
- Random forest can run capably on huge databases.
- Random forest can handle an N quantity of input data without variable removal.
- It provides the most essential features in the classification.
- It can execute well even if the data are omitted.

The limitations of Random Forest are:
- The huge number of trees in random forest can be a reason for interruption in processing.
- It has been noted that random forest is apt only for a few datasets.

6 Proposed Methodology

The proposed method used NSL-KDD data set to calculate the performance of IDS. It is an enhanced version of KDDCup'99 data set. Due to inefficiency in KDD data set revealed by many researchers, we used NSL-KDD data set. It consists of 41 features of a network connection that include training data and

International Journal of Scientific & Engineering Research Volume 7, Issue 10, October-2016
ISSN 2229-5518

testing data. It is significant to select the appropriate attributes from the data set. Here, a proficient algorithm is used for feature selection, i.e. random projection technique. The feature selection techniques are used to choose a subset of appropriate features for the model building. The main profit of using a feature selection procedure consist of the simplification of models that can be easily analyzed by the users or researches; they decrease the number of features and go for only the relevant attributes for the arrangement process which in turn reduce the processing time.

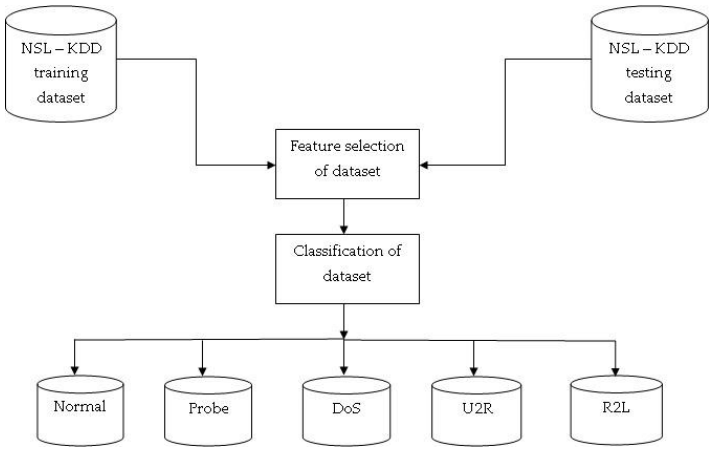

Fig. 3 Proposed Method

Random projection is a way of diminishing the dimensionality of a variety of points. The gain of using random projection task is it is a simple process that can produce fewer invalid result compared with other methods. Dimensionality reduction method reduces the number of variables using various mathematical procedures that have been used in machine learning. Dimensionality reduction is often used to reduce the problem of organization and manipulating huge data sets. Random projection can accomplish sooner processing times and smaller model sizes. The dimensions of random projection matrices are managed so as to approximately preserve the distances between any two samples of the dataset. The main idea of this technique is that the points in a vector space of high dimension can be projected to an apt lower-dimensional space. It can be achieved in a way that considers two points that can preserve distances between the points. It states that the original n-dimensional data is projected to a m-dimensional (m << n) subspace using a random m x n- dimensional matrix A whose rows have unit lengths. Using matrix notation: If n is the original set of N n-dimensional observations, then m is the projection of the data onto a lower m-dimensional subspace. Random projection form a random matrix "A" and project the data matrix N onto M dimensions of order m x n. After that two classification technique SVM and random forest have been used for the classification of data set. The classification process is used to allocate the unknown data into a particular class depending upon predefined data.

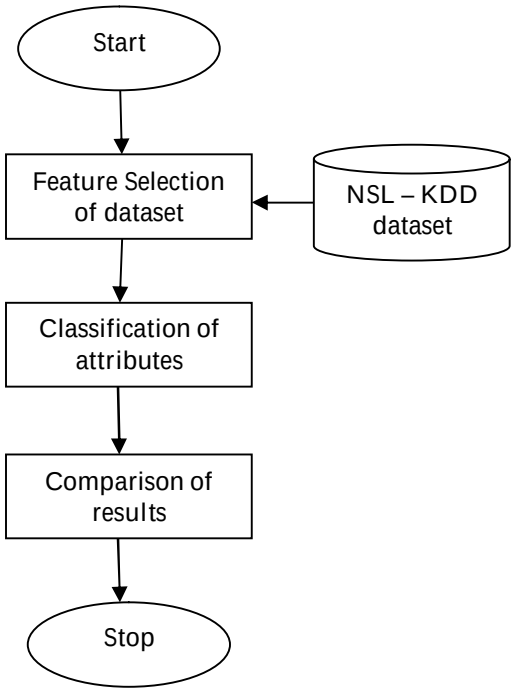

Fig.4 Flowchart of Proposed Method

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. SVM is a supervised learning practice that mainly analyzes the data set or the unknown data for classification. With the help of training samples it sort the upcoming data into various classes. A set of teaching examples are given each marked as either one of the two categories. The SVM training algorithm constructs a model. This model can assign new data into one group or the other. The classification of data can be done by this valuable procedure. The main aim of support vector machine is to obtain a model that cam predicts the end values of the test data. For support vector machine it is compulsory that each data should be a real number in order to represent it in vector space. So, if there are any categorical values then it should be converted into real numbers in order to present it on vector space. There will be a linear separation between the data that has to be classified. With the help of this line the data can be easily separated with more accuracy. In this technique, two categories are available; we can either classify the data to one class or to the other class depending upon the behavior of the new data. The SVMs can also perform non-linear classification. The non – linear classification can be done using kernel that means it can implicitly map their records into high-dimensional spaces. In such a context, unsupervised learning approach is implemented, which attempts to cluster the data with similar behaviors.

7 Experimental Results

This section deals about the experimental setup and result

International Journal of Scientific & Engineering Research Volume 7, Issue 10, October-2016
ISSN 2229-5518

analysis. Many standard data mining process such as data pre-processing, clustering, classification, regression, visualization, feature selection and so on are already implemented in WEKA. We used WEKA 3.8.0 to judge the classifiers such as random forest and support vector machine. The NSL – KDD data set is used to estimate the performance of the IDS. The data set consist of various classes of attacks namely DoS, R2L, U2R and probe. The data set to be classified is initially pre-processed and random projection technique is used as feature selection technique to reduce the dimensionality of the features available in the data set 41 to 11 attribute set. Then classification is done by using SVM and random forest classifiers.

Table 1 Comparison of Proposed Classifiers

Classification Algorithm	Classes	Accuracy (%)
Random Projection + Random Forest	Probe	99.5
	DoS	100
	U2R	75.8
	R2L	98.6
Random Projection + SVM	Probe	99.3
	DoS	99.8
	U2R	0
	R2L	71.8
Random Forest	Probe	98.2
	DoS	99.2
	U2R	86.2
	R2L	54.0
SVM	Probe	70.1
	DoS	96.8
	U2R	15.7
	R2L	2.2

Accuracy = (TP / TP + FP) * 100

From the above classification table, we can understand that the first classifier random forest is better than that of the support vector machine. The classification accuracy for probe is 99.5% for random forest with random projection and the classification accuracy for SVM with random projection is 99.3% which shows the random forest is more efficient than SVM for the classification. The classification of attack classes shows that for classification of probe and DoS good accuracy rate is for random forest than SVM. But, U2R for SVM it is 0% when compared with random forest which means the SVM classifier shows very poor performance for the classification. Overall, we can conclude that random forest is more efficient than SVM.

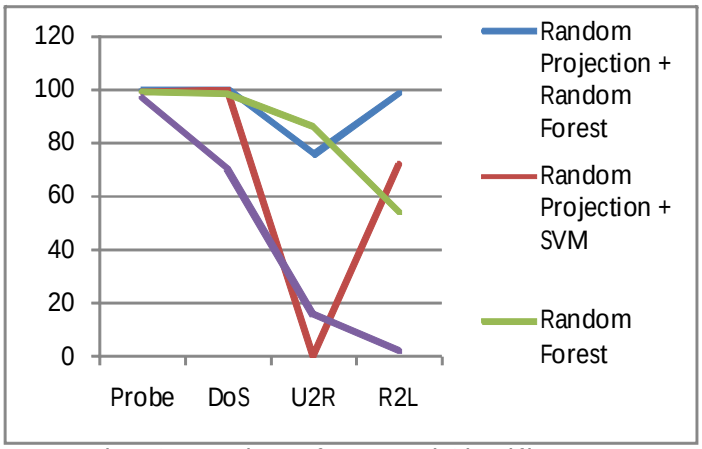

Fig.5 Comparison of Proposed Classifiers

Detection Rate = (TP / TP + FN) * 100

The parameter true positive, false positive, true negative & false negative are used to evaluate the classifiers performance. The detection rate can be calculated by TP and FN. In the above equation TP stands for true positive and FN stands for False Negative. True Positive (TP) is the malicious traffic is correctly identified. False Negative (FN) is the malicious traffic is allowed to exist unchecked. False Positive (FP) is the traffic incorrectly identified as malicious. True Negative (TN) is analyzed as the event that is correctly classified as normal. From the above table, we can see that the detection rate of random forest is more than that of SVM. One of the proposed classifier i.e., random forest achieves approximately 100% detection rate. We can conclude that the classifier random forest is the most efficient algorithm among the various classification techniques.

Table 2 Detection Rate of Proposed Algorithms

Classification Algorithm	Detection Rate (%)
Random Projection + Random Forest	100
Random Projection + SVM	98.5
Random Forest	99.85
SVM	99.52

International Journal of Scientific & Engineering Research Volume 7, Issue 10, October-2016
ISSN 2229-5518

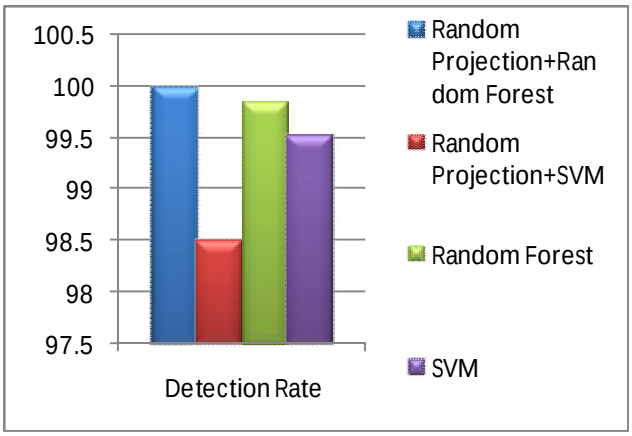

Fig.6 Comparison of Detection Rate of Proposed Classifiers

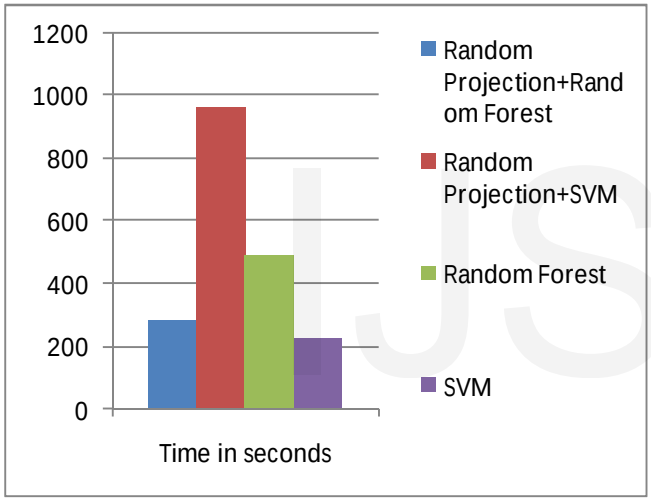

Fig. 7 Comparison of time of proposed classifiers

From the above graph which represents the time comparisons of the proposed algorithms, we can clearly understand that the time taken to build the model for random projection and random forest is 279.61 seconds which is less when compared with random forest without using random projection that obtained the result in 491 seconds. It shows that random projection plus random forest produced better result in less time when compared with random forest without random projection. When we compare the second algorithm random projection plus SVM, it produced the result in 963.9 seconds. It is highest time taken among other proposed algorithms which result in a conclusion that this is the poor classifier among the proposed algorithms. The fourth classifier SVM produced a result 222.28 second which can be considered as the best classifier among other classifiers proposed.

8 Conclusion

In this paper, IDS based on SVM and random forest along with random projection is implemented. The performance of IDS is evaluated with the help of NSL – KDD dataset. The IDS is trained and tested by means of data set. The proposed work used random projection technique one of the efficient feature selection technique to choose the best attributes from the data set. They grouped the features to generate a good yield. The classification techniques SVM; random forest along with random projection is used. The detection rate of random projection along with random forest is approximately 100% and the detection rate of random projection along with SVM is 98.5% which is a good detection rate. But, one factor is time management, as we can see, the time taken for the random projection with random forest to obtain the result is very less than the second classifier i.e. SVM with random projection. From the comparison table, we can conclude that random forest along with random projection yielded the best output than support vector machine along with random projection. The various attack classes such as DoS, probe, U2R and R2L are listed along with their accuracy value. The accuracy rate for DoS is 100% for random projection with random forest. This is an efficient rate when compared with other proposed algorithm. In SVM, the accuracy rate for U2R attack is very poor. The accuracy for probe attack in random projection along with random forest is 99.5% which is a good rate when compared with SVM. Finally, we can conclude that random forest is the best classifiers amidst the other classifiers.

As a future work, the proposed IDS can be implemented on network for protecting it from unlawful activity. In addition, it can be implemented for http services, ftp services for the detection of unauthorized work.

References

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

[2] Meghana Solanki, Vidya Dhamdhere “Intrusion detection by k-means clustering, C 4.5, FNN and SVM classifier. In: International Journal of Emerging Trends & Technology in Computer Science volume 3, Issue 6, November – December 2014.

[3] 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)

[4] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu and Ali. A “ A Detailed Ananlysis of the KDD CUP’99 Data Set”, Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defence Applications (CISDA 2009).

[5] 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.

[6] S. Lee, G. Kim, and S. Kim, Sequence –Order-Independent Network Profiling for Detecting Application Layer DDoS Attacks, EURASIP J. Wireless Comm. and Networking, vol. 2011, no. 1, p. 50, 2011.

[7] The NSL-KDD data set. <http://nsl.cs.unb.ca/NSL-KDD/>

[8] KDD Cup 1999 Data.<http://kdd.ics.uci.edu/databases/kdd-cup99/kddcup99.html>

[9] H. Hajji, “Statistical analysis of network traffic for adaptive faults

International Journal of Scientific & Engineering Research Volume 7, Issue 10, October-2016
ISSN 2229-5518

detection", IEEE Trans. Neural Netw., vol. 16, no. 5, pp. 1053-1063, Sept. 2005.

[10] M. Thottan and C. Ji, "Anomaly detection in IP networks", IEEE Trans. Signal Process, vol. 51, no. 8, pp. 2191-2204, Aug. 2003.

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

[12] G. H. John, R. Kohavi, and K. Pfleger," Irrelevent features and the subset selection problem", in Proc. Int. Conf. Machine Learning, New Brunswick, NJ, USA, July 1994, pp. 121-129.

[13] 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.

[14] Bilal Maqbool Beigh, M. A. Peer,"Intrusion Detection and Prevention System: Classification and Quick Review", ARPN Journal of Science and Technology, vol. 2, No. 7, August 2012.

[15] A. Hussain, J. Heidemann, and C. Papadopoulos," Identification of repeated denial of service attacks", in Proceedings of the 25th IEEE International Conference on Computer Communications. Barcelona, Spain: IEEE, April 2006, pp. 1 – 15.

[16] 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.

[17] 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.