Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 11, Number 6 (2015), pp. 5163-5173

http://www.ripublication.com

A Novel Data Mining Approach for Intrusion Detection System with SVM & GA Model

S.Vijayarangam[1*], A.Rajesh[2]

[1]Research Scholar, Department of Computer Science and Engineering, St.Peter's University, Avadi, Chennai, India, email: skbvijay@yahoo.com*

[2]Professor& Head, Department of Computer Science and Engineering, C.Abdul Hakeem College of Engineering and Technology, Melvisharam, Vellore, India, email: amrajesh73@gmail.com

Abstract

Intrusion detection system is the emergent area of research in the secured computer networks with the advanced usage of the internet in daily life. An Intrusion detection system (IDS) is important in assuring the security for networks and various resources. Recently the data mining techniques have gained importance of network security issues, including network intrusion detection. Hence in this paper, we proposed Intrusion Detection System using data mining technique and classification process will be done by using Support Vector Machine (SVM) algorithm, which may classify the dataset into the different types to identify the attacker packets. In first stage we use a feature selection process with KDD Data set and in second stage we preprocess and classified the data by GA and SVM then finally the data is tested and validated to produce the detection ration, false positive ratio and accuracy. In order to improve the accuracy we used NSL-KDD dataset with a new version of the KDD Cup99 dataset. This type of system is proposed to improve the accuracy of the intrusion detection, when compared to other conventional classifiers which proposed in earlier, such as ANN and KNN classifier. Our proposed result also demonstrates that the accuracy of detecting attacks fairly well.

Keyword : Intrusion Detection System (IDS), NSL-KDD Dataset, Genetic Algorithm (GA), Support Vector Machine (SVM).

Introduction

In today's information most of the computer network is adulterate by various threats such as viruses, trojan horses, worms, intrusions, etc. These viruses can be greatly controlled by installing antivirus software and updating the virus files. In the beginning days researchers studied the various computer security techniques, namely cryptography, firewalls, anomaly and intrusion detection. Among them network intrusion detection is one of the most important technique for defending complex and dynamic intrusion behaviors. In recent years, this method using data mining has attracted more and more interest. Intrusion Detection System (IDS) as an important link in the network security infrastructure is to identify the denial of service attacks, port scans and attempt the network, crack into monitoring network traffic and prevent a future use of known exploits. The IDS annoy with false positive alerts and authorize the security professionals to be affected by the analysis tasks perform [1]. IDS achieved by various techniques in order to improve the probability of detection of suspect threats while decrease the false positive risk.

In the Data mining technique, IDS is one of the areas which are due to limited scalability, adaptability and validity. The Intrusion data is gathered from various sources like network log data, host data etc. In this data analysis is too hard while the network traffic is very large. The use of IDS along with different Data mining techniques for intrusion system of classification and clustering easily excerpt the information from large dataset. In our previous work we are using KDD99 data set to measure the efficiency of the system and obtain the detection rate. To validate the effectiveness and feasibility of the proposed technique, we have used NSL-KDD dataset and it is a new version of KDDcup99. The NSL-KDD dataset has clarified some inherent problems of existing data sets.

In the field of intrusion detection have two different approaches can be noticed. According to the detection approaches are Misuse detection and Anomaly detection [2]. Firstly, the Misuse detection is defined by signature based IDS were found the intrusions are based on the familiar attacks like antivirus software. This antivirus software analyzes the particular data with known code of virus. In Misuse detection, the dataset is used to store the known malicious activity and detect the suspicious data by comparing the stored pattern of attacks. Next the Anomaly detection is also known as behavior detection. It is very different from Misuse detection. Anomaly detection considers that an intrusion will reflect some deviation from normal patterns. These anomaly detectors are divided into static and dynamic detection. The static detector is addressing the software part of the system and based on the assumption that the hardware needs not to be checked. The static detectors targets on integrity checking. A dynamic detector usually operates on audit records or monitored network traffic data. Audit records of systems do not record all events and only recorded in the audit will be noticed then the events may occur in sequence. The partial ordering of events is enough for detection in distributed system.

There are two types of intrusion detection system, namely as Host based intrusion detection and Network-based intrusion detection. The Network based IDS consider the network traffic to observe the threats that produce the abnormal traffic flows, such as a DoS attack, scanning and certain structures of malware. The Host-Based IDS

defined, it audits the characteristics of a single host and the event's occurrance within the host and Figure.1 shows the working Intrusion Detection system. In this paper, we demonstrate Support Vector Machine (SVM) with conventional Genetic Algorithm (GA) to improve the speed of convergence and improve the accuracy.

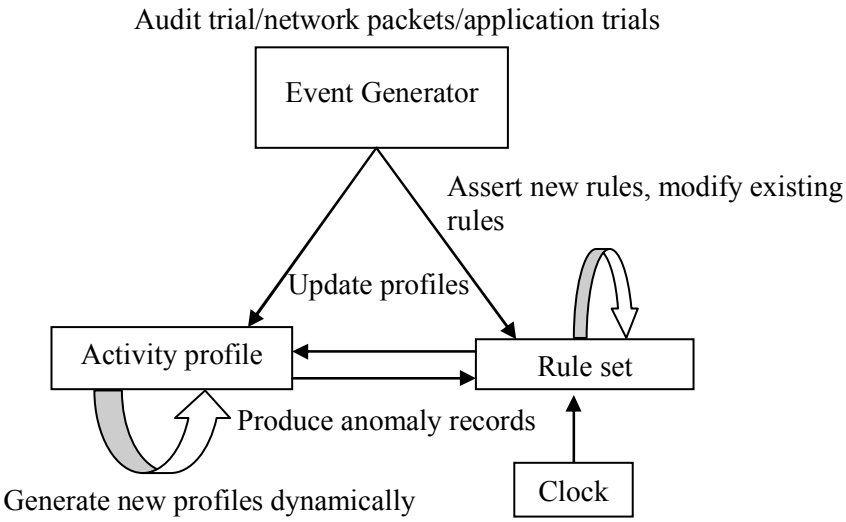

Figure.1. Working of Intrusion Detection

Background Study

In [3] has represented one conspire IDS module to produce a real time detection and block intrusions before occur based on a HIDS using system called anomaly detection. In [4] designed a multiple level tree classifier for intrusion detection system and improve detection rate. In this paper used classifier is more powerful in the case of known attacks, but for unknown attacks gives a low detection rate. In [5] proposed a model intrusion detection joining decision tree and support vector machine techniques and gives high detection rate. In [6] presents the K-nearest neighbor classifier implements on intrusion system and classify the performance in terms of the runtime and error rate on the usual malicious dataset. Proposed new model KDD dataset [7] performed perfectly in terms of accuracy and advantage of this model is the runtime much less as a fewer number of features are used for classification. In [8] mentioned, applying genetic algorithms with fuzzy logic is given for system to efficiently find many types of network intrusions. In this paper they are using the fuzzy confusion matrix where the fuzzy membership value and function for the complement of a fuzzy with different concepts. This system can upload the new rules to the system as the new intrusion detection process. Therefore, it is cost may be very effective and adaptive. The SVM and ANN are used for the detection in system and comparing with each other to produce the effectiveness of the GA [9] on these methods. The GA with ANN classifier requires 18 features and achieves the 98% of accuracy rates. In this paper, they are used KDDcup 99 datasets for detection of five categories of network attacks.

Proposed Approach

Genetic Algorithm (GA) Overview:

GA is a simple method that detects an approximate solution to an optimization task. It is a family of computational methods based on principles of evolution and natural selection, which using a chromosome-like data structure and develops the chromosomes using selection, crossover and mutation operators. A simple genetic algorithm may be the design of a population generator and a selector, a fitness estimator and three estimate operators namely selection, mutation and crossover [10]. The mutation operator is used to convert the particular bits with some probability. The crossover operator combines the two individual species and produces two new off springs. The springs are used to replace low fitness individual in the population [11]. The searching process will be concluded while after giving some number of generations. A genetic algorithm is a programming technique which is estimated as problem-solving strategy. Following the GA operators are applied on a population of chromosomes.

a) Selection: This operator detects which chromosome from the population will be selected for recombination, which is based on the fitness of the chromosomes. These kinds of chromosomes are called parents. The selection methods are namely as: fitness, proportion selections, roulette-wheel selection, stochastic universal sampling, local selection and rank selection.

b) Crossover: The crossover methods are used to recombine the parent chromosomes. It will be produced new chromosomes are called offspring. The crossover methods are single point crossover, multipoint crossover, uniform and arithmetic crossover.

c) Mutation: The mutation process is used to introduce the new genetic material into the new population process. This process will improve the diversity in the population.

The advantages of Genetic algorithm are mentioned as below:

(i) The process of genetic algorithm is parallel. Because of multiple offspring can be produced which is utilized as a solution.

(ii) It suggests different solutions for some problem.

(iii) Also develop the new rules or increase the set of intrusion detection.

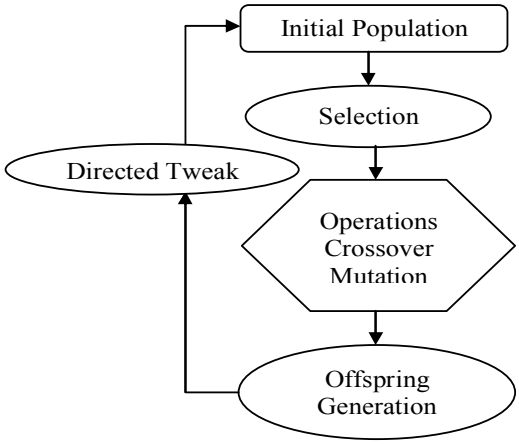

Figure.2 Flow diagram of Genetic Algorithm

Proposed Operation:

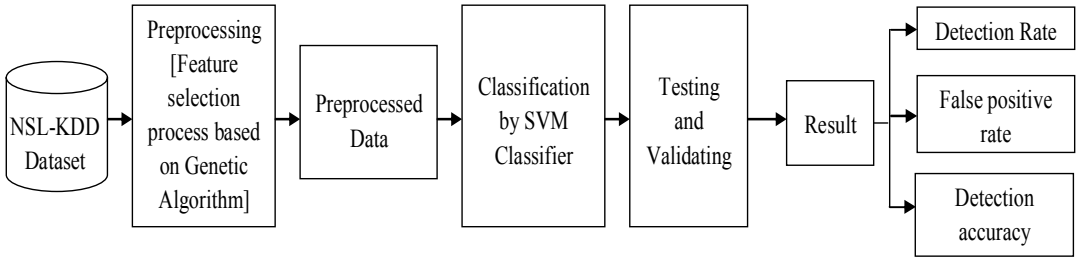

Figure.3 Block diagram of the proposed method

1. Data Set Collection:

The collection of the dataset is used to verify the productivity and feasibility of the Proposed Intrusion Detection System. In this method we have used NSL-KDD dataset. It is an advanced version of KDDcup99 dataset [12]. But our previous work we are used KDDcup99 dataset. The NSL-KDD dataset has resolved the some inherent problems of the KDDcup99 dataset, which may be recognized as a standard benchmark for intrusion detection evaluation. The NSL-KDD dataset is designed approximately 4,900,000 single connection vectors and each vector can be contained 41 features. This type of new dataset is similar to NSL-KDD dataset and design as normal or attack type.

The below reasons are mentioned as NSL-KDD dataset has become more popular dataset than KDDcup99 dataset for IDS purpose.

(a) It is used to reject the redundant records from the training set.
(b) Duplicate records can be removed from the training set and to increase the intrusion detection performance.
(c) The NSL-KDD dataset provides the accurate determination of different learning techniques.
(d) Our dataset is inexpensive to use for experimental purpose as it designed of reasonable numbers of examples both in the training and testing set.

2. Preprocessing Stage:

2.1. Feature Selection process based on Genetic algorithm:

The preprocessing was needed before the SVM classifier system could be designed. The preprocessing was consisted of symbol valued attributes to numeric, scaling and attack names. The preprocessing includes the following processes:

- Mapped the symbolic features to a numeric value.
- The scaling values have been implemented since for the data significantly changing the resolution and ranges. The data are scaled to fall within the range [-1, 1]
- The attacks are generalized to one of the five classes, Normal Attack, Denial of Service (DOS), User to Root (U to R), Remote to Local (R to L), Probe.

Generally the final training set is the output of the data preprocessing that extract the knowledge for the testing phase. The data preprocessing includes the learning, normalization, transformation, transformation, feature extraction and feature selection process. In our work, feature selection process based on genetic algorithms for preprocessing stage [13]. The feature selection of the subset was depended on the accuracy of the classifier. In this method used for choosing the subset of features from the initial data set. In existing method, the feature selection process contains filter method and wrapped method. The filter method based on the data features, characteristics without containing the machine language. The main advantage of filter method is low computational cost and without affecting the machine language algorithm for feature selection. The wrapper method is mainly used for feature subset selection from the dataset based on functional analysis of the feature dataset. But in our proposed method, the Genetic Algorithm (GA) is used to select a feature dataset for a preprocessing stage. This algorithm can be decrease the PMU feature from 41 attributes to 6 attributes are similar to the characteristics of DoS attack which may reduce 85% of the feature space. The six attributes are namely as, Protocol, src_bytes, dst_bytes, count, srv_count, same_srvrate. Our NSL-KDD dataset is containing an enormous number of redundancy records. The dataset contains 10% full DoS attack and 71% testing phase has fully affected the decision of IDS. The algorithm of feature selection based on GA is mentioned below:

Algorithm	
Step 1:	Initialize the pre-processed data
Step 2:	Determine the DoS attack for individual preprocessed data.
Step 3:	Choose the individual selection
Step 4:	Perform the operation of mutation and pair of individuals
Step 5:	Evaluate the objective function for created population
Step 6:	If (5) step is satisfied, conclude the operation and if not satisfied repeat the step 3.
Step 7:	Recover the best feature from the dataset which may be reflects the DoS properties.

3. Support Vector Machine (SVM) Classifier:

In the data mining area, the support vector machine is one of the most successful classification algorithms. The SVM is used to determine a hyperplane to operate binary classification and SVM used as a high dimensional space. The SVM classifier is based on the idea of hyper plane classifier and this classification technique is based on Statistical Learning Theory (SLT). The objective of SVM is to detect a linear optimal hyperplane so that margin of separation between the two classes is maximized. The SVM use a part of the data to train the system and classifier is represents the training data. It is a valuable technique due to its high accuracy and performance in solving regression and classification tasks. The training time in SVM is an expensive task as a full time used to solve a problem.

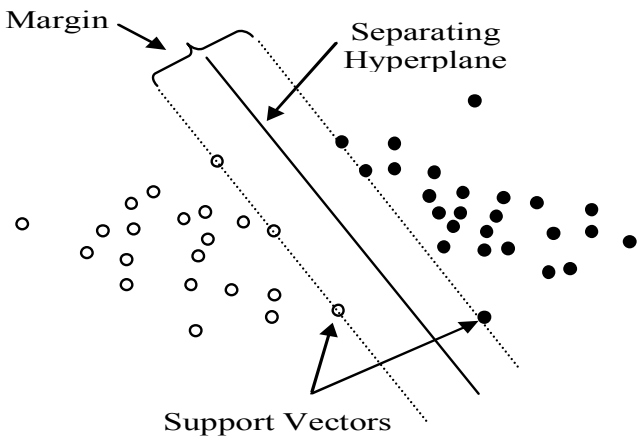

Figure.4 separate between the hyper planes with SVM

The SVM is introduced from some principle which is namely as Structural Risk Minimization (SRM) principle. SVM is mainly disturbed with classes and disconnect the data in a hyperplane defined by a number of support vectors [14] [15] in Figure.3. This type of support vectors are subdivided into training data used to define the boundary between the two classes. If we assume the SVM cannot support to separate data in hyperplane and the kernel function is used to project the data into high-dimensional feature space. Figure.4 represents the SVM algorithm and high dimensional feature space is to produce hyperplane which is allowing a linear separation. The kernel function is the very valuable in Support Vector Machine (SVM) and used to determine the support vectors. The kernel functions are three type's namely as linear, polynomial or Gaussian.

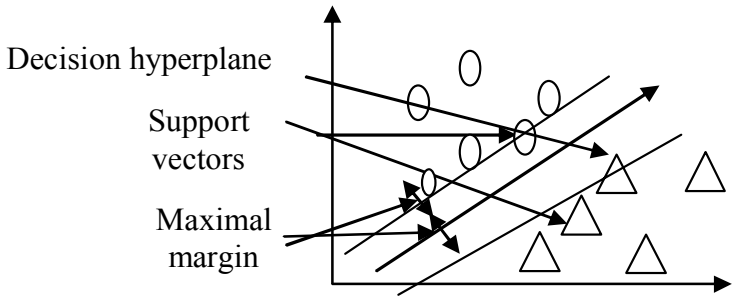

Figure.5 Support Vector Machine

Separate the number of training vectors are separate two classes to be considered as a problem of SVM classification. Where the two classes are (x_1, y_1), (x_2, y_2),……. (x_n, y_n), the corresponding class labels are mentioned as $x_i \in R^p$ And $y_i \in \{-1, +1\}\}$, $1 \leq i \leq n$. The $f(x,\theta)$ is a main function of classifier and it can be determined. Such that

$y=f(x,\theta)$, Where y is mentioned as the label of class for x, θ is a vector of function of unknown parameters. Generally the SVM model algorithm describes the maximal margin in hyperplanes into separate two classes, which are needed to solve the following problem can be defined as,

$$\boldsymbol{minimize_{w,b}\frac{1}{2}w^2} \quad (1)$$

To subject the optimal hyperplane:

$$\boldsymbol{y_i(w^T x_i + b) \geq 1, i = 1, 2, 3, \ldots\ldots, n} \quad (2)$$

$$\boldsymbol{\xi_i \geq 0, i = 1, 2, 3, \ldots\ldots, n} \quad (3)$$

To acknowledge the errors, the optimization problems are mentioned as:

$$\boldsymbol{min_{w,b}\frac{1}{2}w^2 + C\sum_{i=1}^{n}\xi_i}$$

Using the Lagrange multipliers method, we evaluate the dual formulation which is declare in terms of variables a_i:

$$\boldsymbol{maximize_a \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j}$$

Subject to: $\boldsymbol{\sum_{i=1}^{n} y_i\alpha_i = 0, 0 < \alpha_i < |C}$ for all $\boldsymbol{i = 1, 2, 3, \ldots..n}$

Finally, the equation (4) denoted as a linear discriminant function based on a linear classifier,

$$\boldsymbol{f(x) = \sum_i^n \alpha_i x_i^T x + b} \quad (4)$$

The non-linear classifier gives better accuracy for many applications. The discriminant function can be mapped on the input space X to a feature space using a non-linear Ø: X→F. We denoted space F, the kernel function can be taken as following terms:

$$\boldsymbol{maximize_a \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k|(x_i, x_j)}$$

Where,

$$\boldsymbol{\sum_{i=1}^{n} y_i\alpha_i = 0, 0 < \alpha_i < C} \quad (5)$$

For all $\boldsymbol{i = 1,,,3,\ldots,,}$

$$\boldsymbol{f(x) = \sum_{i}^{n} \alpha_i k\,(x, x_i) + b}$$

The above equation is mentioned as a discriminant function of the kernel function. The SVM classifications are formulated for two class problems. Many types of support vector machines are used to handle the multiclass problems.

Experimental Results

A set of data is elected to train the process and the algorithm. Figure.5 represents the block diagram of our proposed method. The set of data is collected from the NSL-KDD dataset, then applied the preprocessing step. We have selected genetic algorithm for feature selection techniques, and processed the data. Thus we process

classification technique using SVM classifier into attack and normal packets. Tested the results and calculate the detection rate, false positive rate and detection accuracy. Also check the performance of the proposed model for the intrusion detection and classification method, we can evaluate the NSL-KDD datasets these five attacks are namely as Normal, Dos, U2R, R2L and probe.

Where,

FN is False negative
TN is True Negative
TP is True positive
FP is False positive

Table.1

Types of traffic	Intrusion detection using SVM algorithm			
	True positive Rate	**True Negative Rate**	**False positive rate**	**False negative rate**
Normal	0.700	0.947	0.0.17	0.300
Dos	0.996	0.702	0.293	0.012
R2L	0.640	0.873	0.035	0.306
U2R	0.658	0.978	0.001	0.371
Probe	0.916	0.897	0.014	0.321

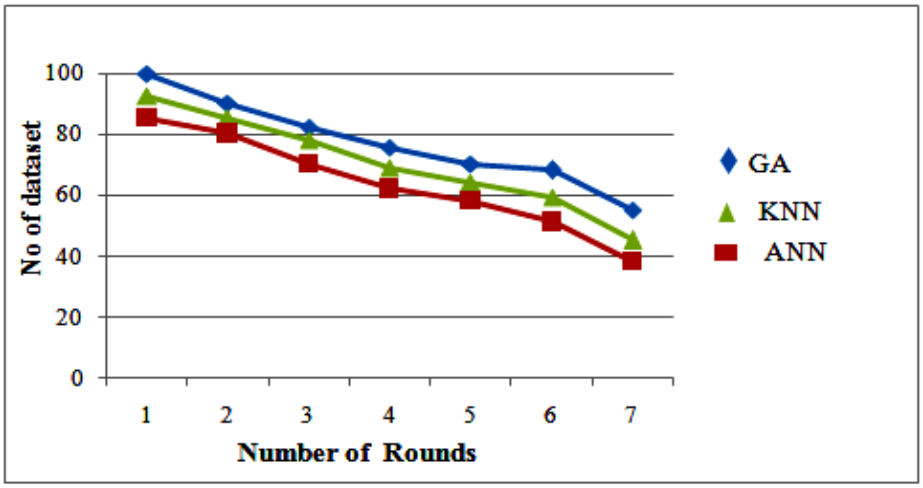

Figure.6 shows the accuracy performance of our method

In Figure ,6 the proposed method is compared with other techniques such as ANN algorithm [16] and KNN algorithm [17]. The overall performance of our proposed approach can provide a high intrusion detection in various rounds with improved accuracy than the other approaches for highly secured communication.

Conclusion

In this research work, we developed an intrusion Detection process is to detect the intrusion into a computer system in order to improve the security and accuracy. The intrusion detection is an area in which more sensitive data are stored and processed in a network system. In this paper, we have proposed a method of intrusion detection using SVM classification which can decrease the time required to design a model for classification and improve the accuracy when we used feature selection process. The Feature selection process is based on genetic algorithm is used to reduce the classification time and store the memory space effectively. We consider that there are many techniques which give good detection rate, system accuracy in case of Denial of Service (DoS) attack, U2R, R2L and probing method.

References

[1] A.A.Ojugo, A.O. Eboka, O.E.Okonta, R.E Yoro, F.O.Aghware. "Genetic Algorithm Rule-based Intrusion Detection System (GAIDS)". Journal of Emerging Trends in computing and Information System (GAIDS). Vol.3, No.8. Aug 2012.

[2] Yunlu Gong. Intrusion detection system combining misuse detection and anomaly detection using Genetic Network Programming. IEEE. Pp- 3463 – 3467. 2009

[3] Kaining Lu Zehua Chen Zhigang Jin Jichang Guo. "An Adaptive Real-Time Intrusion Detection System Using Sequences of System Call", Journal of Computer security. 2003.

[4] Xiang, M.Y. Chong and H. L. Zhu, "Design of Multiple-level Tree classifiers for intrusion detection system", IEEE conference on Cybernetics and Intelligent system, 2004.

[5] Peddabachigiri S., A. Abraham., C. Grosan and J. Thomas, "Modeling of Intrusion Detection System Using Hybrid intelligent systems", Journals of network computer application, 2007.

[6] M.Govindarajan and Rlvl.Chandrasekaran, "Intrusion Detection Using k-Nearest Neighbor" pp 13-20, ICAC, IEEE, 2009

[7] JingTao Yao, Songlun Zhao, and Lisa Fan. An Enhanced Support Vector Machine Model for Intrusion Detection. http://kdd.ics.uci.edu/databases/ kddcup99/kddcup99.html

[8] Wenke Lee, Salvatore J. Stolfo, o K. "A Data Mining Framework for Building Intrusion Detection Models". Security and Privacy, Proceedings of the 1999 IEEE Symposium on. Pp-120 – 132. 1999.

[9] Amin Dastanpour, Suhaimi Ibrahim, Reza Mashinchi, Ali Selamat. "Comparison of Genetic Algorithm Optimization on Artificial Neural Network and Support Vector Machine in Intrusion Detection System". 2014 IEEE Conference on Open Systems (ICOS), October 26-28, 2014, Subang, Malaysia.

[10] Shaveta , Er. Abhinav Bhandari and Dr. Krishan Kumar Saluja. "Applying Genetic Algorithm in Intrusion Detection System: A Comprehensive Review". Proc. of Int. Conf. on Recent Trends in Information, Telecommunication and Computing, ITC,ACEEE. 2014.

[11] Ren Hui Gong, Mohammad Zulkernine, Purang Abolmaesumi. "A Software Implementation of a Genetic Algorithm Based Approach to Network Intrusion Detection". Proceedings of the Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks (SNPD/SAWN'05). 2005 IEEE.

[12] Srinivas Mukkamala, Guadalupe Janoski, "Andrew Sung. Intrusion Detection Using Neural Networks and Support Vector Machines". IEEE, 2002.

[13] Dharmendra G. Bhatti, P. V. Virparia. "Data Preprocessing for Reducing False Positive Rate in Intrusion Detection". International Journal of Computer Applications (0975 – 8887). Volume 57– No.5, November 2012.

[14] G.V. Nadiammai, M. Hemalatha. "Effective approach toward Intrusion Detection System using data mining techniques". Production and hosting by Elsevier. Egyptian Informatics Journal. (2014) 15, 37–50.

[15] Mrs. Snehal A. Mulay. Prof. P. R. Devale, Prof. G.V. Garje. "Decision Tree based Support Vector Machine for Intrusion Detection". 20 10 International Conference on Networking and Information Technology, IEEE. 2010

[16] M.Govindarajan, Rlvl.Chandrasekaran. "Intrusion Detection Using k-Nearest Neighbor". 2009 IEEE.

[17] Wun-Hwa Chen, Sheng-Hsun Hsu∗, Hwang-Pin Shen. "Application of SVM and ANN for intrusion detection". Computers & Operations Research, ELSEVIER. 32 (2005) 2617 – 2634.