International Journal of Emerging Engineering Research and Technology

Volume 2, Issue 2, May 2014, PP 112-118

An Approach for Building Intrusion Detection System by Using Data Mining Techniques

Praveen P Naik
PG Student, Department of CS&E, AIT,
Chikmagalur, Karnataka, India

Prashantha S J
Asst Professor, Department of CS&E, AIT,
Chikmagalur, Karnataka, India

Abstract: *Information security is one of the cornerstones of Information Society. Integrity and privacy of financial transactions, personal information and critical infrastructure data, all depend on the availability of strong and trustworthy security mechanisms. In recent years, many researchers are using data mining techniques for building IDS. Here, we propose a new approach by utilizing data mining techniques such as neuro-fuzzy and radial basis support vector machine (SVM) for helping IDS to attain higher detection rate. The proposed technique has four major steps: primarily, k-means clustering is used to generate different training subsets. Then, based on the obtained training subsets, different neuro-fuzzy models are trained. Subsequently, a vector for SVM classification is formed and in the end, classification using radial SVM is performed to detect intrusion has happened or not. To illustrate the applicability and capability of the new approach, the results of experiments on KDD CUP 1999 dataset is demonstrated. Experimental results shows that our proposed new approach do better than Conditional random fields (CRF) with respect to specificity and detection accuracy.*

Keywords: *Intrusion Detection System (IDS), K-Means, Neuro-fuzzy, SVM, CRF, Data Mining.*

1. INTRODUCTION

An Intrusion Detection System (IDS) is a device (or application) that monitors network and/or system activities for malicious activities or policy violations and produces reports to a Management Station. Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies, acceptable use policies, or standard security practices.

Intrusion Detection Systems have undergone rapid growth in power, scope and complexity in their short history. In recent years, Intrusion detection system has been one of the most sought after research topics in the field of Information Security having huge applications in the cooperate world where data integrity and security is a complex issue.

When an intruder attempts to break into an information system or performs an action not legally allowed, we refer to this activity as an Intrusion. Intruders may be external or internal depending upon the authorization level. Intrusion techniques may include exploiting software bugs or system configurations, password cracking, sniffing unsecured traffic, or exploiting the design flaw of specific protocols.

An Intrusion Detection System (IDS) is a system for detecting intrusions and reporting them accurately to the proper authority. IDSs are usually specific to the operating system that they operate in and are an important tool in the overall implementation of an organization's information security policy, which reflects an organization's statement by defining the rules and practices to provide security, handle intrusions, and recover from damage caused by security breaches.

2. LITERATURE SURVEY

In this section, related literature about machine learning approach and preparation of datasets for data mining activity will be reviewed and discussed.

Annie George [3], Anomaly detection has emerged as an important technique in many application areas mainly for network security. Anomaly detection based on machine learning algorithms considered as the classification problem on the network data has been presented here. Dimensionality reduction and classification algorithms are explored and evaluated using KDD99 dataset for network IDS. Principal Component Analysis for dimensionality reduction and Support Vector Machine for classification have been considered for the application on network data and the results are analyzed. The result shows the

decrease in execution time for the classification as they reduce the dimension of the input data and also the precision and recall parameter values of the classification algorithm shows that the SVM with PCA method is more accurate as the number of misclassification decreases.

W.K. Lee, S.J.Stolfo [4], there is often the need to update an installed intrusion detection system (IDS) due to new attack methods or upgraded computing environments. Since many current IDSs are constructed by manual encoding of expert knowledge, changes to IDSs are expensive and slow. This paper describes a data mining framework for adaptively building Intrusion Detection (ID) models. The central idea is to utilize auditing programs to extract an extensive set of features that describe each network connection or host session, and apply data mining programs to learn rules that accurately capture the behavior of intrusions and normal activities. These rules can then be used for misuse detection and anomaly detection. New detection models are incorporated into an existing IDS through a meta-learning (or co-operative learning) process, which produces a meta detection model that combines evidence from multiple models. We discuss the strengths of our data mining programs, namely, classification, meta-learning, association rules, and frequent episodes. We report on the results of applying these programs to the extensively gathered network audit data for the 1998 DARPA Intrusion Detection Evaluation Program

V. Jyothsna, V. V. Rama Prasad, K. Munivara Prasad [5], With the advent of anomaly-based intrusion detection systems, many approaches and techniques have been developed to track novel attacks on the systems. High detection rate of 98% at a low alarm rate of 1% can be achieved by using these techniques. Though anomaly-based approaches are efficient, signature-based detection is preferred for mainstream implementation of intrusion detection systems. As a variety of anomaly detection techniques were suggested, it is difficult to compare the strengths, weaknesses of these methods. The reason why industries don't favor the anomaly-based intrusion detection methods can be well understood by validating the efficiencies of the all the methods. To investigate this issue, the current state of the experiment practice in the field of anomaly-based intrusion detection is reviewed and survey recent studies in this. This paper contains summarization study and identification of the drawbacks of formerly surveyed works.

CHEN Bo, Ma Wu [6], the effective way of improving the efficiency of intrusion detection is to reduce the heavy data process workload. In this paper, the dimensionality reduction use of technology in the classic dimensionality reduction algorithm principal component to analysis large-scale data source for reduced-made features of the original data be retained and improved the efficiency of intrusion detection. And use BP neural network training the data after dimensionality reduction, will be effective in normal and abnormal data distinction, and achieved good results.

Paul Dokas , Vipin kumar [7], in which they gives an overview of our research in building rare class prediction models for identifying known intrusions and their variations and anomaly/outlier detection schemes for detecting novel attacks whose nature is unknown. Disadvantage of this paper is that due to the fact that the number of instances of U2R and R2L attacks in the training data set is very low, these numbers are not adequate as a standard performance measure. It could be biased if we use these numbers as a measure for performance of the system.

3. METHODOLOGY

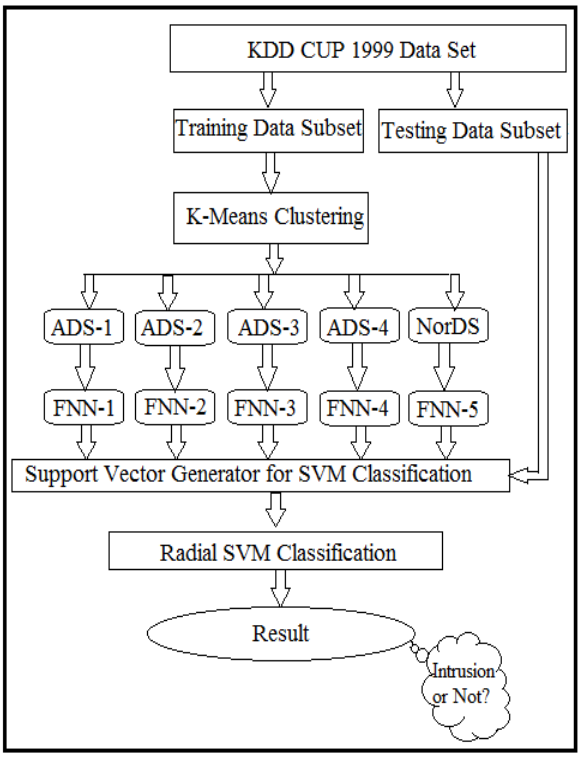

Fig 1. *Architecture for Proposed IDS*

The system architecture of proposed technique as shown in the fig 1 consists of 5 step methodology. This can be explained as follows.

1. The input data set DS needed for experimentation is prepared by conducting relevance analysis on KDD Cup 1999 data set in order to reduce the irrelevant attributes / features which will not contribute for intrusion detection.
2. The input dataset is divided into Training Data set and Testing Data set. The Training data is clustered using K-Means Clustering into k subsets where, k is the number of clusters desired.
3. Neuro-fuzzy (FNN) training is given to each of the k cluster, where each of the data in a particular cluster is trained with the respective neural network associated with each of the cluster.
4. Generation of vector for SVM classification, $S=\{D_1, D_2,\ldots\ldots D_N\}$ which consists of attribute values obtained by passing each of the data through all of the trained Neuro-fuzzy classifiers, and an additional attribute μ_{ij} which has membership value of each of the data.
5. Classification using SVM to detect intrusion has happened or not.

The detailed description of each of the steps is elaborated in the following sub-sections.

3.1 Data Collection

This section, it gives an overview of the data set used for intrusion detection. This data set contains seven weeks of training data and two weeks of testing data. The raw data was about four gigabytes of compressed binary TCP dump data from the of network traffic generated. This was processed into about five million connection records, each

of which is a vector of extracted feature values of that network connection. As we know, a connection is a sequence of TCP packets to and from some IP addresses, starting and ending at some well defined times. This data set of the five million connection records was used as the data set for the 1999 KDD intrusion detection contest and is called the KDD Cup 99 data. In particular, MIT Lincoln Lab's DARPA intrusion detection evaluation datasets have been employed to design and test intrusion detection systems. In 1999, recorded network traffic from the DARPA 98 Lincoln Lab dataset was summarized into network connections with 41-features per connection. This formed the KDD 99 intrusion detection benchmark in the International Knowledge Discovery and Data Mining Tools Competition.

The KDD 99 intrusion detection datasets are based on the 1998 DARPA initiative, which provides designers of intrusion detection systems (IDS) with a benchmark on which to evaluate different methodologies [3]. To do so, a simulation is made of a factitious military network consisting of three 'target' machines running various operating systems and services. Additional three machines are then used to spoof different IP addresses to generate traffic. Finally, there is a sniffer that records all network traffic using the TCP dump format. The total simulated period is seven weeks. Each connection was labeled as normal or as exactly one specific kind of attack. All labels are assumed to be correct.

There were a total of 37 attack types in the data set. The simulated attacks fell in exactly one of the four categories : User to Root; Remote to Local; Denial of Service; and Probe.

- **Denial of Service (dos):** Attacker tries to prevent legitimate users from using a service.
- **Remote to Local (r2l):** Attacker does not have an account on the victim machine, hence tries to gain access.
- **User to Root (u2r):** Attacker has local access to the victim machine and tries to gain super user privileges.
- **Probe:** Attacker tries to gain information about the target host.

Data preprocessing comprises following components including document conversion, feature selection and feature weighting.

The functionality of each component is described as follows:

[1] Dataset prepared with DOS attack which include smurf, Neptune, back, teardrop and POD ping of death attacks /anomaly.

[2] Feature selection – reduces the dimensionality of the data space by removing irrelevant or less relevant feature selection criterion.

[3] Document conversion- converts different types of documents such as gz, tcpdump to csv file and arff (Attribute-Relation File Format) data file format.

[4] Totally we considered 11850 data points for our experimentation.

3.2 K-Means Clustering

K-means is one of the simplest unsupervised learning algorithms that solve the clustering problem. The objective is to classify a given data set into a certain number of clusters (assume initial clusters) fixed a priori.

The pseudo code for the **adapted K-Mean algorithm** is presented as below

1. Choose random k data points as initial Clusters Mean (Cluster center)

2. Repeat

3. for each data point x from D

4. Computer the distance x and each cluster mean (centroid)

5. Assign x to the nearest cluster.

6. End for

7. Re-compute the mean for current cluster collections.

8. Until reaching stable cluster

9. Use these centroid for normal and anomaly traffic.

10. Calculate distance of centroid from normal and anomaly centroid points.

11. If distance(X, Dj) > = 5

12. Then anomaly found ; exit

13. Else then

14. X is normal;

The k-means clustering algorithm is based on finding data clusters in a data set by keeping minimized cost function of dissimilarity measure. In most cases this dissimilarity measure is chosen as the Euclidean distance. For each data point to be clustered, the cluster centroid with the minimal Euclidean distance from the data point will be the cluster for which the data point will be a member.

$$j = \sum_{j=1}^{K} \sum_{i=1}^{n} \left\| x_i^{(j)} - C_j \right\|^2$$

3.3 Artificial Neural Network

ANN is a biologically inspired form of distributed Computation. It is composed of simple processing units, or nodes, and connections between them. The connection between any two units has some weight, which is used to determine how much one unit will affect the other. A subset of the units acts as Input nodes and another subset acts as output nodes, which perform summation and threshold. The ANN has successfully been applied in different fields. The feed-forward neural network trained with the back-propagation algorithm is a common tool for intrusion detection.

ANN module aims to learn the pattern of every subset. ANN is a biologically inspired form of distributed computation. It is composed of simple processing units, and connections between them. In this study, we will employ classic feed-forward neural networks trained with the back-propagation algorithm to predict intrusion.

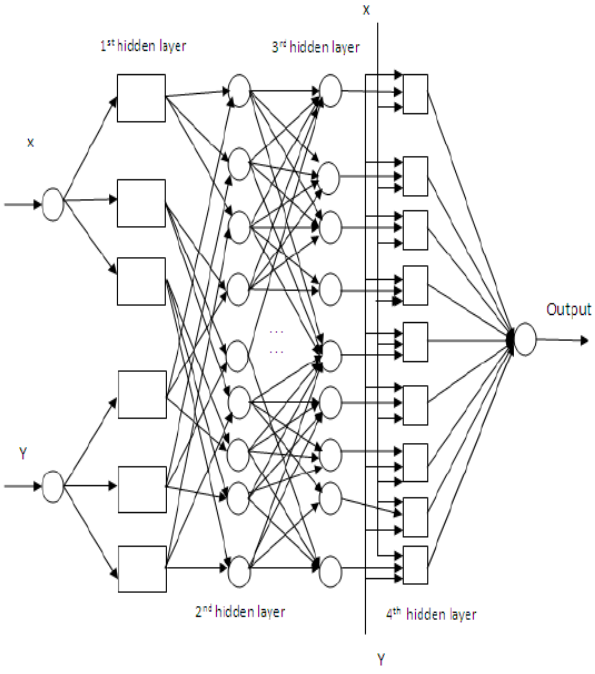

Fig .2. *Neuro-fuzzy architecture*

A feed-forward neural network has an input layer, an output layer, with one or more hidden layers in between the input and output layer. The ANN functions as follows: each node i in the input layer has a signal xi as network's input, multiplied by a weight.

Classification of the data point considering all its attributes is a very difficult task and takes much time for the processing, hence decreasing the number of attributes related with each of the data point is of paramount importance. Executing the reduced amount of data also results in decrease of error rate and the improved performance of the classifier system.

The main purpose of the proposed technique is to decrease the number of attributes associated

with each data, so that classification can be made in a simpler and easier way. Neuro-fuzzy classifier is employed to efficiently decrease the number of attributes.

3.4 Radial SVM Classification

In our system, we are employing radial SVM for the final classification for the intrusion Detection. SVM is used as it achieves enhanced results when contrasted to other classification Techniques especially when it comes to binary classification. In the final classification, the data is binary classified to detect intrusion or not.

The input data is trained with neuro-fuzzy after the initial clustering as we have discussed earlier, then the vector necessary for the SVM is generated. Here in the process, each of the data is fed into each of the neural classifier to get the output value. That is each of the data is fed into K number of neuro-fuzzy classifiers to yield K output values. So the data values gets distorted and after passing through the K neuro-fuzzy classifiers, attribute number of the data in consideration changes and diminishes to K numbers where each value will be the output of the data passing through the respective neuro-fuzzy.

The vector array S= $\{D_1, D_2, ...D_N\}$ where, D_i is the i^{th} data and 'N' is a total number of input data. Here, after training through the neuro-fuzzy the attribute number reduces to 'k' numbers. $D_i = \{a_1,a_2,.....a_k\}$, here the D_i is the i^{th} data governed by attribute values a_i, where a_i will have the value after passing through the i^{th} neuro-fuzzy. Total number of neuro-fuzzy classifiers trained will be K, corresponding to the K clusters formed after clustering. we comprise a parameter known as membership value. Inclusion of the membership value into the attribute list results in a better performance of the classifier. Membership value μ_{ij} is defined by the equation below.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_i\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

Hence, the SVM vector is modified as $S^*=\{D^*_1, D^*_2,.....D^*_N\}$ where S^* is the modified SVM vector which consists of modified data D^*_I, which consists of an extra attribute of membership value μ_{ij}.

$D^*_1 = \{a_1,a_2,.....a_k, \mu_{ij}\}$, Hence the attribute number is reduced to K+1 where K is the number of clusters. This results in simple processing in the final SVM classification. This is due to the fact that input data which had 34 attributes is now constrained to K+1 i.e. to 6 attributes. This also reduces the system complexity and time incurred.

Use of radial SVM results in obtaining better results from the classification process when compared to normal linear SVM. In linear SVM, the classification is made by use of linear hyper-planes where as in radial SVM, nonlinear kernel functions are used and the resulting maximum-margin hyper-plane fits in a transformed feature space. The corresponding feature space is a Hilbert space of infinite dimensions, when the kernel used is a Gaussian radial basis function. The Gaussian Radial Basics function is given by the equation:

$$\phi(x - x_j) = \exp\left(-\frac{1}{2\sigma_j^2} \|x - x_j\|^2 \right)$$

Where j=1,2,….N. The 'j'th input data point x_j defines the center of radial basis function, the vector 'x' is the pattern applied to the input. σ_j is a measure of width of 'j'th Guassian function with center x_j.

4. RESULTS AND DISCUSSION

A. Screen Shots and Output

This snap shot (fig 3) shows IDS mining form in which user has to open KDD 99 Data set. When the user clicked on perform k-means button then clusters will be formed displaying table including ID, attributes, and types of training data points.

There will be totally 5 number of clusters namely PROBE,DOS(denial of service),R2L(remote to local),U2R(user to root),Normal.

After formation of clusters, training and classification of data points is easy and less time consuming and less complexity. Next we have to perform neuro-fuzzy operation to train our training data. fig 4 shows the neuro-fuzzy form.

We consider only seven attributes namely service, login access, no. of times, dst_bytes, connection status, duration of connection and src_bytes. Classification of the data point considering all its attributes is a very difficult task and takes much time for the processing,

hence decreasing the number of attributes related with each of the data point is important.

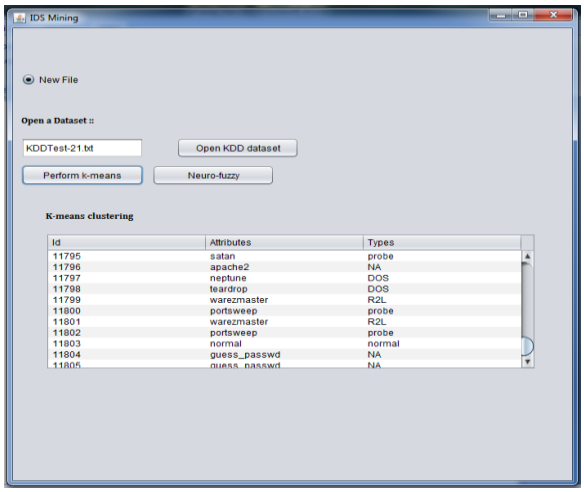

Fig 3. *Screen shot for IDS Mining form*

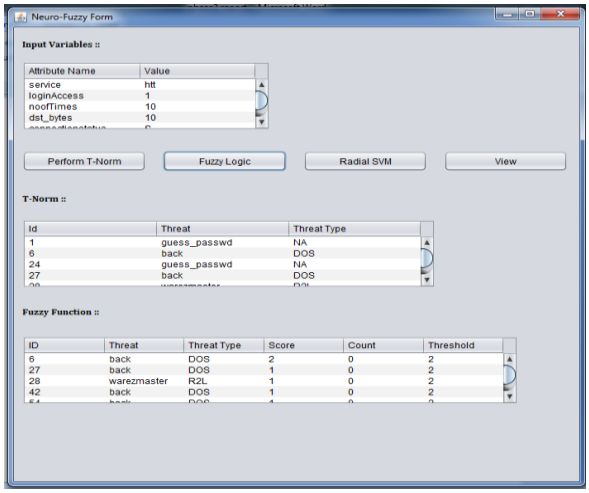

Fig 4.. *Screen shot for neuro-fuzzy form*

After performing normalization operation, the next step is to perform neuro-fuzzy operation as shown in the above fig 3. In which we consider score, count and threshold values obtained after normalization of the data points.

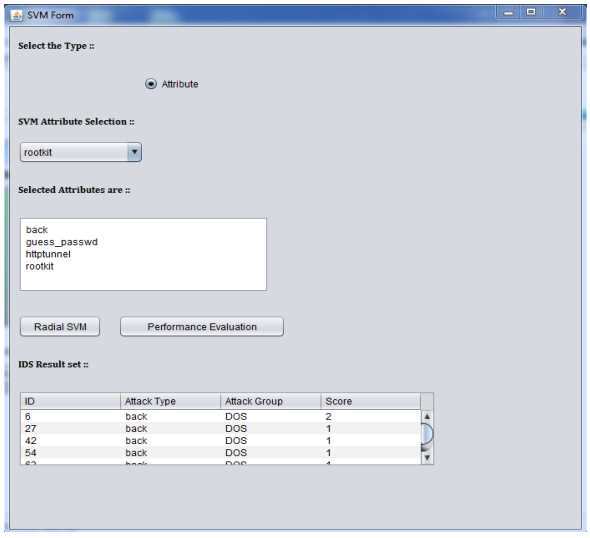

Fig 5. *Screen shot for SVM Classification form*

At the end, radial SVM classification is performed to detect intrusion happened or not. This phase includes SVM attribute selection step. At the end we will get SVM Result set, which shows intrusions if happened with their ID, Attack types, Attack group and score.

The measurement used for evaluation of our proposed techniques are True positive (TP), False negative (FN), True negative (TN), and False positive (FP).

- **True Positive**- A legitimate attack which triggers IDS to produce an alarm.
- **False Positive**- An event signaling IDS to produce an alarm when no attack has taken place.
- **False Negative**- A failure of IDS to detect an actual attack.
- **True Negative**- When no attack has taken place and no alarm is raised.

The graphical representation of comparison of Conditional Random Fields (CRF) with our proposed Technique is shown below in fig 5.

This shows our proposed technique for building IDS by using data mining techniques such as k-means clustering, Neuro-fuzzy training and radial support vector machine (SVM) do better than CRF in terms of performance.

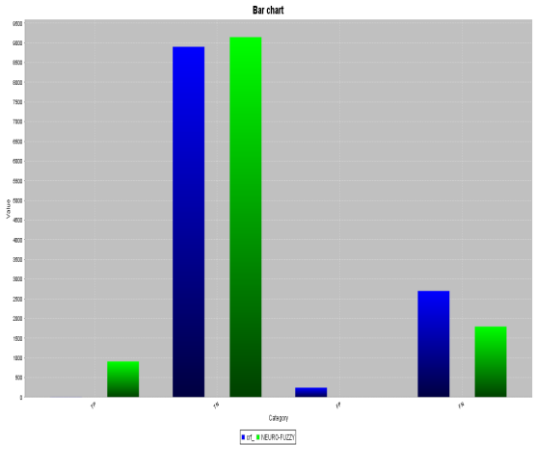

Fig 6. *Performance comparison chart for SVM and CRF*

5. CONCLUSION

In recent years, research on neural network methods and machine learning techniques to improve the network security by examining the behavior of the network as well as that of threats is done in the rapid force. The large volume of database is increasing rapidly resulting in gradual rise in the security attacks. The current IDS is ineffective to update the audit data rapidly it involves human interference thus

reduces the performances. The paper elaborates the architecture of the Intrusion Detection System along with features of an ideal intrusion detection system. The study also describes the categorization and challenges if the IDS. In this paper we analyzed the neural network approach and the machine learning approach in overcoming the challenges of the IDS. Further there is need to design the system which will overcome the current challenges of IDS and also the system must provide a high performance in detecting the threats and security attacks. Presently the application support only 10% KDD CUP dataset. This application can be extended to manage more number of records.

REFERENCES

[1] A.M.Chandrasekhar and K.Raghuveer, "Intrusion Detection Technique by using K-means, Fuzzy Neural Network and SVM classifiers", presented at International Conference on Computer Communication and Informatics (ICCCI-2013), Coimbatore, INDIA.

[2] Sandip Ashok Shivarkar, and Mininath Raosaheb Bendre, "Hybrid Approach for Intrusion Detection Using Conditional Random Fields", International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 1, Issue 3.

[3] Annie George, "Anomaly Detection based on Machine Learning: Dimensionality Reduction using PCA and Classification using SVM", International Journal of Computer Applications (0975 – 8887) Volume 47– No.21, June 2012.

[4] W.K. Lee, S.J.Stolfo, "A data mining framework for building intrusion detection model", In: Gong L., Reiter M.K. (eds.): Proceedings of the IEEE Symposium on Security and Privacy. Oakland, CA: IEEE Computer Society Press, pp.120~132, 1999.

[5] V. Jyothsna, V. V. Rama Prasad, K. Munivara Prasad, "A Review of Anomaly based Intrusion Detection Systems", International Journal of Computer Applications (0975 – 8887) Volume 28– No.7, August 2011.

[6] CHEN Bo, Ma Wu, ―Research of Intrusion Detection based on Principal Components Analysis‖, Information Engineering Institute, Dalian University, China, Second International Conference on Information and Computing Science, 2009.

[7] Paul Dokas , Vipin kumar, "Data Mining for Network Intrusion Detection", Proceeding of NGDM., pp.21-30, 2002.

[8] A.M.Chandrashekhar and K. Raghuveer. "Performance evaluation of data clustering techniques using KDD Cup-99 Intrusion detection data set " International Journal of Information & Network Security (IJINS), Vol.1, No.4, pp.294-305, 2012.

[9] Jose Vieira, Fernando Morgado Dias, Alexandre Mota. "Neuro-Fuzzy Systems: A Survey". Proceeding Internal Conference on Neural Networks and Applications, 2004.

[10] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", Proceeding IEEE international conference on computational intelligence for security and defence applications, pp. 53-58, Ottawa, Ontario, Canada, 2009.