International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-3 Issue-5 September-2012

Design Hybrid method for intrusion detection using Ensemble cluster classification and SOM network

Deepak Rathore[1], Anurag Jain[2]
Department of Computer Science & Engineering, RITS, Bhopal[1]

Abstract

In current scenario of internet technology security is big challenge. Internet network threats by various cyber-attack and loss the system data and degrade the performance of host computer. In this sense intrusion detection are challenging field of research in concern of network security based on firewall and some rule based detection technique. In this paper we proposed an Ensemble Cluster Classification technique using som network for detection of mixed variable data generated by malicious software for attack purpose in host system. In our methodology SOM network control the iteration of distance of different parameters of ensembling our experimental result show that better empirical evaluation on KDD data set 99 in comparison of existing ensemble classifier.

Keywords

Intrusion Detection, ECC, SOM, KDD-CUP 99

1. Introduction

In the past years, there were few intruders and so the user can carry off them easily from the known or unknown attacks. Intrusion detection attacks can be classified into two groups: Misuse or Signature based and Anomaly based Intrusion Detection. The misuse or signature based intrusion detection system discovers the intrusion by equating with its existing signatures in the database. The detecting attacks and signatures are matching, it's an intrusion. The signature based intrusions are called known attacks whenever the users are detecting the intrusion by coping with the signatures log files. The log file comprises the listing of known attacks detecting from the computer system or networks. The anomaly based intrusion detection is called as unknown attacks and this attack is observed from network as it departs from the normal attacks.

The network based attacks are detected from the interconnection of computer systems. The system is able to communicate with each other, so that the attack is sent from one computer system to another computer system by the way of routers and switches. The host based approaches are detected only from a single computer system and is easy to prevent the attacks. These attacks mainly come about from some external devices which are connected. The external devices are pen drive, CD, VCD, Floppy and DVD etc. The web based attacks are likely when systems are connected over the internet and the attacks can be spread into different systems through the e-mail, chatting, downloading the materials etc. Nowadays many computer systems are struck from web based dangerous attacks. An intrusion can be determined as "any set of actions that attempts to compromise the integrity, confidentiality, or availability of a resource". User authentication (e.g., using passwords or biometrics), avoiding programming errors, and information protection (e.g., encryption) have all been used to protect computer systems. As systems become more composite. Elements central to intrusion detection are resources to be protected in a target system, i.e., user accounts, file systems, system kernels, etc.; models that characterize the "normal" or "legitimate" behaviour of these resources; techniques that compare the actual system activities with the established models identifying those that are "abnormal" or "intrusive". In pursuit of a secure system, different measures of system CONDUCT have been proposed, on the basis of an ad hoc presumption that normalcy and anomaly (or illegitimacy) will be accurately manifested in the chosen set of system features.

Intrusion detection system deal with supervising the incidents happening in computer system or network environments and examining them for signs of possible events, which are infringement or imminent threats to computer security, or standard security practices Intrusion detection systems (IDS) have emerged to detect actions which endanger the integrity, confidentiality or availability of are source as an effort to provide a solution to existing security issues. This technology is relatively new, however, since its beginnings, an enormous number of proposals have been put forward to sort this situation

out in the most efficient and cost effective of manners.

The Self-Organizing Map is a neural network model for analysing and visualizing high dimensional data. It belongs to the category of competitive learning network. The SOM defines a mapping from high dimensional input data space onto a regular two-dimensional array of in designed architecture is input vector with six input values and output is realized to 2 dimension spaces.

The SOM is a neural network trained with a competitive learning rule in an unsupervised manner. A competitive learning rule means that the neurons compete to respond to a stimulus, such as a connection vector (recall that a connection vector describes properties of a network connection, such as the destination port and number of packets sent).The neuron that is most excited by the stimulus, i.e. whose weight vector is most similar to the connection vector, wins the competition. The winning neuron earns the right to respond to that stimulus in future, and the learning rule adjusts its weight vector so that its response to that stimulus in future will be enhanced, i.e. by moving the weight vector closer to the connection vector. This means that the next time that same connection vector is presented, the neuron that won the competition for that same vector last time will be more excited by it. During training, the SOM learns to project connection vectors that are close together (in terms of Euclidean distance) onto neurons that are close to each in the output grid. In this way, the SOM learns relationships between the connections a vector, expressing them as spatial relationships in the output grid. The training algorithm also ensures that the weight vectors of the neurons area good representation of the connection vectors in the training data. This is achieved by aiming for a low mean quantisation error, where the quantisation error is the distance between a connection vector and the winning neuron's weight vector. The mean quantisation error is the average of this over all connection vectors in the training set.

KDD99 Dataset CUP:
The KDD99 dataset is now the benchmark for training, testing and evaluating learning IDS,so it is basic for IDS developers. The competition task was to build a network intrusion detector, a predictive model or a classifier that cantle what are "bad" connections, called intrusions or attacks, and what are "good", called normal connections These attacks fall into four main categories:

DOS: denial-of-service, e.g. syn flood.
R2L: unauthorized access from a remote machine, e.g. guessing password;
U2R: unauthorized access to local super user (root) privileges, e.g., various "buffer overflow" attacks;
Probing: surveillance and other probing, e.g., port scanning

2. Related Work

The signature based intrusion is detected using neural network classifier like Feed Forward Neural Network (FFNN), Probabilistic Neural Network (PNN) And Radial Basis Neural Network (RBNN). The various techniques are applied in this problem in MATLAB application for improving the best performance applied to KDD Cup 1999 dataset. The performance of the full featured dataset and reduced dataset is analysed[1].The classifications of intrusion detection and methods of data mining applied on them were introduced. Then, intrusion detection system design and implementation of based on data mining were presented. Such a system used. APRIORI algorithm to analyse data association, which is the most influencing algorithm in mining Boolean association rules continuity item muster, with recurrence arithmetic based on idea of two period continuity item muster as core. Experiments showed that new type of attack can be detected effectively in the system, and knowledge base can be updated automatically, so the efficiency and accuracy of the intrusion detection were improved, and security of the network was enhanced[5]. Network security is becoming an increasingly important issue, since the rapid development of the Internet. Network Intrusion Detection system (IDS), as the main security defending technique, is widely used against such malicious attacks. Data mining and machine learning technology has been extensively applied in network intrusion detection and prevention systems by discovering user behaviour patterns from the network traffic data. Association rules and sequence rules are the main technique of data mining for intrusion detection .Considering the classical Apriori algorithm with bottleneck of frequent item sets mining, we propose a Length-Decreasing Support to detect intrusion based on data mining, which is an improved Apriori algorithm. Experiment results indicate that the proposed method is efficient [12].

Intrusion detection systems aim to identify attacks with a high detection rate and a low false alarm rate.

Classification-based data mining models for intrusion detection are often ineffective in dealing with dynamic changes in intrusion patterns and characteristics. Consequently, unsupervised learning methods have been given a closer look for network intrusion detection. Traditional instance-based learning method scan only be used to detect known intrusions, since these methods classify instances based on what they have learned. They rarely detect new intrusions since these intrusion classes has not been able to detect new intrusions as well as known intrusions. In this paper, we propose a soft Computing technique such as Self-organizing map for detecting the intrusion in network intrusion detection. Problems with k-mean clustering are hard cluster to class assignment, class dominance, and null class problems. The network traffic datasets provided by the NSL-KDD Data set in intrusion detection system which demonstrates the feasibility and promise of unsupervised learning methods for network intrusion detection [11].

A technique of combining K-Means clustering and genetic algorithm to IDS. The training data has been clustered in to 2-clusters before feeding the initial population hoping that data will be divided into normal and abnormal clusters. There were no declared experiment results but the author concluded that his approach detected known and unknown and the results were not good for some runs [8].

A Dependable Network Intrusion Detection System (DNIDS) based on the Combined Strangeness and Isolation measure K-Nearest Neighbor (CSIKNN) algorithm. The intrusion detection algorithm analyzes different characteristics of network data by employing two measures: strangeness and isolation. But in general the K-NN still needs intensive computations. The Unsupervised Anomaly Detection Using an Optimized K-Nearest Neighbors Algorithm can work without the need for massive sets of pre-labeled training data. The author discussed the creation of such a system that uses a k-nearest neighbor's algorithm to detect anomalies in network connections, as well as the optimization necessary to make the algorithm feasible for a real-world system. The drawback of this approach is that the detection rates and false positive rates were not good as other approaches [9].

The Intrusion Detection techniques are used to detect the intrusions based on the KDD Cup 1999 dataset. These dataset contains 41 features in various types of attacks. By reducing 41 features into 13 features the accuracy has improved by 97.5% using the Probabilistic Neural Network. Principal Component Analysis is one of the most widely used dimensionality reduction techniques for data analysis and compression. The Principal Component Analysis is applied to the KDD CUP 1999 dataset to reduce its features and implemented using MAT LAB software. PCA selects 13 features from 41 feature data set. The reduced features are used as input to different classifiers and the results are compared. The results show the efficiency with 13 features is comparable to the 41 features, with reduced training and testing times. Comparing these three classifiers PNN gives better efficiency than FFNN and RBN. The KDD Cup 1999 reduced dataset obtained with PCA shows promising results. Hence, The KDD dataset consists of network connection records generated by a TCP/IP dump. It contains 4, 940, 000 connection records. There are 41 features in each record. 10% of the original data are training data with a label which identifies which category the record belongs. A false negative occurs when an intrusion action has occurred but the system considers it as a non-intrusive behavior. A false positive occurs when the system classifies an action as an intrusion whiles it is a legitimate action. A good intrusion detection system should perform with a high precision and a high recall, as well as a lower false positive rate and a lower false negative rate. To consider both the precision and false negative rate is very important as the normal data usually significantly outnumbers the intrusion data in practice. To only measure the precision of a system is misleading in such a situation. A poor intrusion detection system may have a high precision but a high false negative rate [7].

This paper presents a data mining algorithm based on supervised clustering to learn data patterns and use these patterns for data classification. This algorithm enables a scalable incremental learning of patterns from data with both numeric and nominal variables. Two different methods of combining numeric and nominal variables in calculating the distance between clusters are investigated. The algorithm and test its performance on a number of data sets from various application domains. The prediction accuracy and reliability of the algorithm are analyzed, tested, and compared with those of several other data mining algorithms.

In this study, we extend a scalable, incremental, and supervised clustering and classification algorithm—

CCAS into ECCAS that has the capacity of handling data with both numeric and nominal variables. Two different methods of handling mixed data types are developed. The two methods of ECCAS are tested and compared on a data set with mixed Variable types for intrusion detection. Both methods produce comparable performance to that of the winning algorithm in a Data mining contest on the same data set. The performance on different data sets shows the reliability of ECCAS. The testing results for one data set also show that the five phases of ECCAS reduces the impact of the data presentation order on the prediction accuracy. The number of Grid intervals shows the impact on the prediction accuracy of ECCAS. The ECCAS algorithm and the distance Measure could be used in common data mining applications. We are developing methods to adaptively and dynamically adjust the parameters during training, including the grid-interval configuration and the threshold-controlling outlier removal [3].

Intrusion detection is a software application that monitors network and/or system activities for malicious activities or policy violations and produces reports to a Management Station. Security is becoming big issue for all networks. Hackers and intruders have made many successful attempts to bring down high profile company networks and web services. Intrusion Detection System (IDS) is an important detection that is used as a countermeasure to preserve data integrity and system availability from attacks. The work is implemented in two phases; in first phase clustering by K-means is done and in next step of classification is done with k-nearest neighbors and decision trees. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the interclass similarity. This paper proposes an approach which makes the clusters of similar attacks and in next step of classification with K nearest neighbors and Decision trees it detect the attack types. This method is advantageous over single classifier as it detect better class than single classifier system [10].

3. Methodology

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. Although classification is an effective means for distinguishing groups or classes of objects, it requires the often costly collection and labelling of a large set of training tuples or patterns which the classifier uses to model each group. It is often more desirable to proceed in the reverse direction: First partition the set of data into groups based on data similarity (e.g., using clustering), and then assign labels to the relatively small number of groups. Additional advantages of such a clustering-based process are that it is adaptable to changes and helps single out useful features that distinguish different groups. A very simple classifier can be based on a nearest-neighbour approach. In this method, one simply finds in the *N*-dimensional feature space the closest object from the training set to an object being classified. Since the neighbour is nearby, it is likely to be similar to the object being classified and so is likely to be the same class as that object. Nearest neighbour methods have the advantage that they are easy to implement. They can also give quite good results if the features are chosen carefully (and if they are weighted carefully in the computation of the distance.) [9].

The K-nearest-neighbour (KNN) algorithm measures the distance between a query scenario and a set of scenarios in the data set. Because the distance between two scenarios is dependant of the intervals, it is recommended that resulting distances be scaled such that the arithmetic mean across the dataset is 0 and the standard deviation 1. This can be accomplished by replacing the scalars with according to the following function:

ECC Algorithm:
This paper presents an extended version of CCAS clustering and classification algorithm—supervised (CCAS), (ECCAS) that enables the handling of mixed data types. The application of ECCAS to computer intrusion detection, using network traffic data with mixed data type. We apply ECCAS to intrusion detection using the Knowledge Discovery and Data Mining (KDD) Cup 1999 data [3].

1. Divide dataset into chunk D1, D2, D3, ….Dn+1.
2. Generate discrete random number of seed for generating of cluster.
3. Initialized distance weight factor.
4. Calculate min of data chunk and standard deviation.
5. Compare value at min with seed value.
6. Then generate cluster.

7. Set label of class C1, C2, C3.
8. Assigned training at data.
9. Generate classifier-Merge set of cluster & classifier with label.
10. Calculate standard deviation (error).
11. Ensemble class.

This algorithm simply tells that initially there is a large data. Firstly it is divided in to the small size chunks. Then we us a random number generator which generate a random number is time. This random number works as a seed for each iteration. It means in each iteration we take a new random number. This iterative process generates a set of seed which is used to generate a clusters, i.e. each seed is works as a representative for each cluster.

4. Proposed Work

ECC- SOM:
Extended version of CCAS clustering and classification algorithm supervised (CCAS), with Self-Organizing Map (SOM).

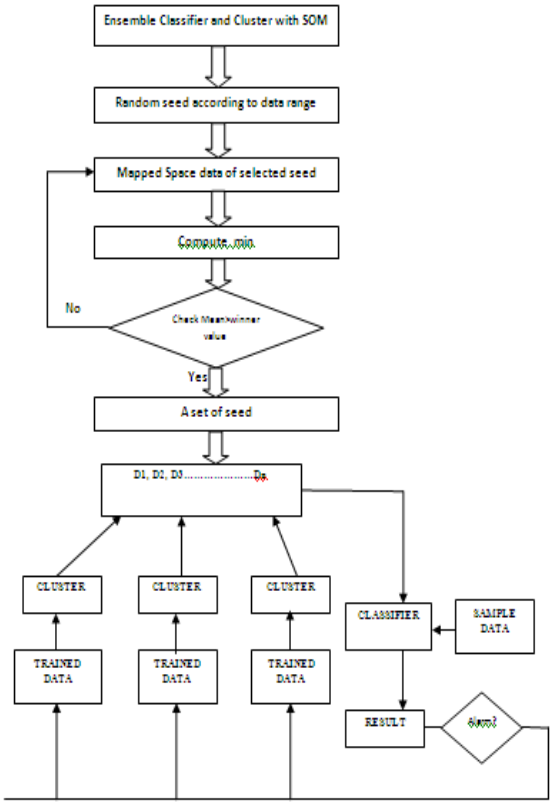

Figure 1: ECC-SOM Model

In this phase of algorithm SOM network apply on ECCA technique. In this process SOM control the iteration of selected data for clustering.

1) Divide dataset into chunk D1, D2, D3, _____Dn+1.
2) Generate discrete random number of seed for generating of cluster.
3) Initialized distance weight factor.
4) Map data into SOM space
5) Calculate min of data chunk and standard deviation.
6) Calculate Winner matrix
7) Compare value at min with seed value.
8) Repeat iteration
9) Then generate cluster.
10) Set label of class C1, C2, C3.
11) Assigned training at data.
12) Generate classifier-Merge set of cluster & classifier with label.
13) Calculate standard deviation (error).
14) Ensemble class.
15) Data classified

5. Result

The figure shown below shows the classification:

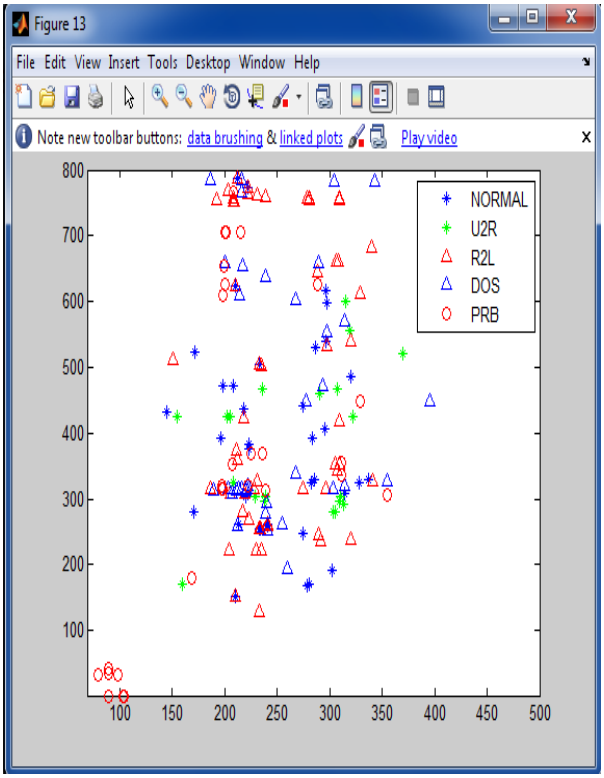

Fig 2: ECC Results

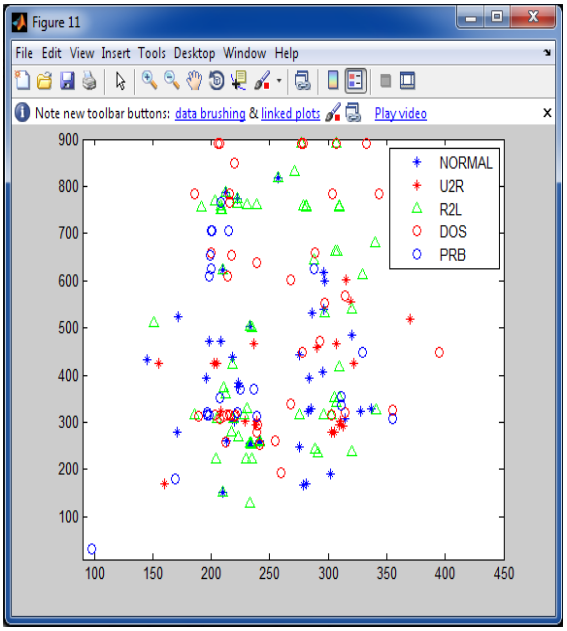

Fig 3: ECC SOM Results

The comparative results are shown in the table drawn below:

Table 1: Comparative Result

Metric		Accuracy (%)	Precision (%)	Recall (%)
Data-Set 1	ECC	92.14	87.24	84.43
	ECC-SOM	97.14	96.11	94.1
Data-Set 2	ECC	89.9	84.32	83.23
	ECC-SOM	95.23	92.14	91.21
Data-Set 3	ECC	91.34	86.14	85.11
	ECC-SOM	95.12	93.21	91.13
Data-Set 4	ECC	92.22	88.21	87.66
	ECC-SOM	97.13	94.52	93.67

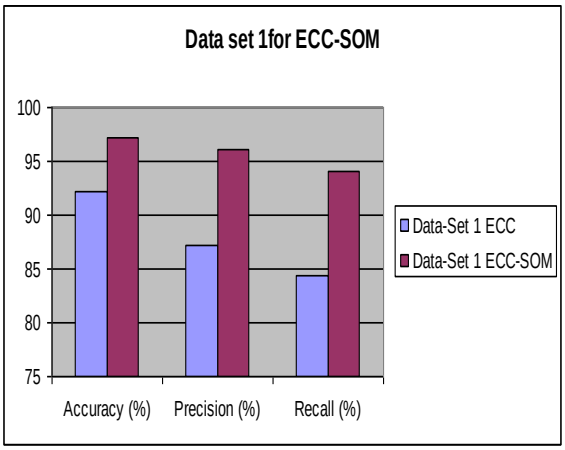

Fig 4: Result of ECC-SOM with Data Set1

Description:- This figure show the Accuracy, Precision, and Recall on ECC and ECC-SOM for data set 1(KDD CUP 99).

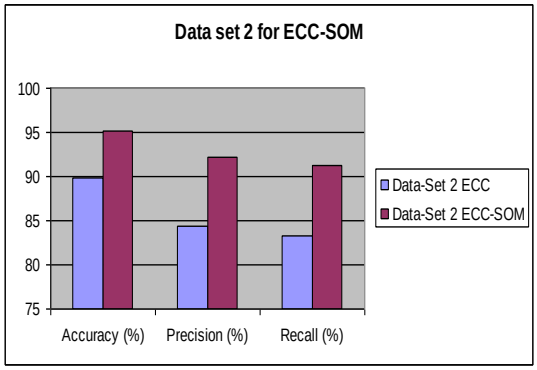

Fig 5: Result for ECC-SOM with Data Set2

Description:- This figure show the Accuracy, Precision, and Recall on ECC and ECC-SOM for data set 2(KDD CUP 99).

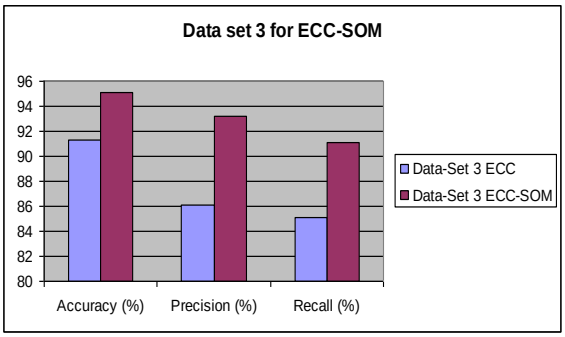

Fig 6: Result for ECC-SOM with Data Set3

Description:- This figure show the Accuracy, Precision, and Recall on ECC and ECC-SOM for data set 3(KDD CUP 99).

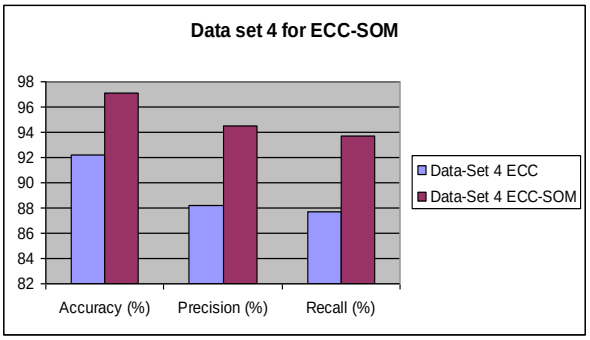

Fig 7: Result for ECC-SOM with Data Set4

Description:- This figure show the Accuracy, Precision, and Recall on ECC and ECC-SOM for data set 4(KDD CUP 99).

6. Conclusion and Future Work

In this paper we proposed ECC-SOM method for intrusion detection on mixed data types. Analysis of result gives a better prediction of result for different data set in KDD, but also suffered problem in Alarm generation. The value of negative alarm is increases and also the value of false positive is increases. Now in future we used trained neural network such as RBF network for the reduction of both false positive rate and false negative rate.

References

[1] S.Devaraju, Dr.S.Ramakrishnan: "analysis of intrusion detection system using various neural network classifiers"1033-1038, IEEE 2011.

[2] K kr.Gupta, B Nath,R Kotagiri" Layered Approach Using Conditional Random Fields for Intrusion Detection" 1545-5971/10/$26.00 © 2010 IEEE.

[3] Li, Xiangyang, and Nong Ye. "A supervised clustering and classification algorithm for mining data with mixed variables." Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 36, no. 2 (2006): 396-406.

[4] Hu, Jiankun, Xinghuo Yu, Dong Qiu, and Hsiao-Hwa Chen. "A simple and efficient hidden Markov model scheme for host-based anomaly intrusion detection." Network, IEEE 23, no. 1 (2009): 42-47.

[5] Miao, Chunyu, and Wei Chen. "A study of intrusion detection system based on data mining." In Information Theory and Information Security (ICITIS), 2010 IEEE International Conference on, pp. 186-189. IEEE, 2010.

[6] Ye, Nong, Syed Masum Emran, Qiang Chen, and Sean Vilbert. "Multivariate statistical analysis of audit trails for host-based intrusion detection." Computers, IEEE Transactions on 51, no. 7 (2002): 810-820.

[7] Tavallaee, Mahbod, Ebrahim Bagheri, Wei Lu, and Ali-A. Ghorbani. "A detailed analysis of the KDD CUP 99 data set." In Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009.

[8] Marimuthu, A., and A. Shanmugam. "Intelligent Progression for anomaly Intrusion detection." In Applied Machine Intelligence and Informatics, 2008. SAMI 2008. 6th International Symposium on, pp. 261-265. IEEE, 2008.

[9] Kuang, Liwei. "DNIDS: a dependable network intrusion detection system using the CSI-KNN algorithm." (2007).

[10] P J. Pathak ,S S. Dongre 2012 **"**Attack Detection By Clustering And Classification Approach" ISSN: 2277–9043 IJAR in Computer Science and Electronics Engineering.

[11] R Ranjani Singh .N Gupta" To Reduce the False Alarm in Intrusion Detection System using self-Organizing Map" ISSN 2250 - 3765 IJCSA .

[12] Lei Li, De-Zhang Yang, Fang-Cheng Shen "A Novel rule-based Intrusion Detection System by"2010.