# THE COMBINED APPROACH FOR ANOMALY DETECTION USING NEURAL NETWORKS AND CLUSTERING TECHNIQUES

A.S. Aneetha and Dr. S. Bose

Department of Computer Science & Engineering,
Anna University, Chennai – 600025, India
asaneetha@annauniv.edu,
sbs@cs.annauniv.edu

## ABSTRACT

*Nowadays detection of new threats has become a necessity for secured communication to provide absolute data confidentiality, integrity and availability. Design and development of such an intrusion detection system in the communication world, should not only be new, accurate and fast but also effective in an environment encompassing the surrounding network. In this paper, a new approach is proposed for network anomaly detection by combining neural network and clustering algorithms. We propose modified Self Organizing Map algorithms which initially starts with null network and grows with the original data space as initial weight vector, updating neighbourhood rules and learning rate dynamically in order to overcome the fixed architecture and random weight vector assignment of simple SOM. New nodes are created using distance threshold parameter and their neighbourhood is identified using connection strength and its learning rule and the weight vector updation is carried out for neighbourhood nodes. The k-means clustering algorithm is employed for grouping similar nodes of Modified SOM into k clusters using similarity measures. Performance of the new approach is evaluated with standard bench mark dataset. The new approach is evaluated using performance metrics such as detection rate and false alarm rate. The result is compared with other individual neural network methods, which shows considerable increase in the detection rate and 2% false alarm rate.*

## KEY WORDS

*Anomaly Detection, Modified SOM, k-means, Weight Vector, Neighbourhood Function*

## 1. INTRODUCTION

In the increased network communication world, security place most important role. One way of providing security is intrusion detection system (IDS), whose basic function is to detect inappropriate, inaccurate and anomalous activity in a system. Attacks may be categories in to any one of these forms namely Denial Of Service (DOS), root to user (R2L), user to root (U2R) and Probing. In communication networks, intrusion detection may be based on network and/or host or based on the application depending on their mode of deployment and data used for analysis. For a network environment, two types of intrusion detection systems, namely, misuse detection or signature based and anomaly detection are very often employed ( Gaddam et al., 2007, Denning

DOI : 10.5121/cseij.2012.2404

1987, Anderson 1980). While the former is capable of identifying known attack patterns with high detection rate and limitation of unable to identify novel attacks, the latter develops a model only with the normal behavior and any deviation from it is classified as an anomaly and therefore new attacks can be identified easily although at the expense of detection rates.

In anomaly detection, machine learning techniques such as classification, clustering and neural network based algorithms (Yasami et al., 2010, Sandhya et al., 2005) are deployed. Most of these techniques, however, work in a supervised environment because inherently they need labeled data. But in a real time environment, these techniques may not be effective as only raw data are available. Therefore, for real time environment, the unsupervised anomaly detection will be more appropriate and efficient, offering many advantages for intrusion detection process. With appropriate modifications, some of the neural network algorithms for unsupervised anomaly detection have been found to be more effective.

Self Organizing Map (SOM) has been reported to be a useful intrusion detection technique for unsupervised learning (Ozgur et al., 2005, Zhi – song et al., 2003, Villamann et al., 1997). SOM has been used to map multi dimensional nonlinear statistical data into two dimensional data space as output. The main set back of this technique, however, is that the number of output nodes is predefined and only the adjacent nodes are taken as neighbourhood. The combination of different unsupervised approaches such neural network and clustering techniques has been reported to be more efficient in anomaly detection (S. Bose et al.2012, Seungmin Lee et al. 2011). In this paper an approach has been made by combining modified SOM and k-means algorithms for anomaly detection. In the modified SOM the drawback of the SOM are rectified by allowing the network to grow, with a distance threshold, and also by using the connection strength to identify the neighbourhood nodes. In k-means the nodes created in the modified SOM are grouped into k clusters using distance measures with their weight vector values as seed points.

## 2. RELATED WORK

Anomaly detection has become an important area of intensive research for secured communication. Many authors have suggested various approaches for unsupervised anomaly intrusion detection with artificial neural networks. In a framework that combined neural network with K-means clustering for the detection of real time anomalies, Seungmin Lee et al. [2011] have reported that new attacks can also be detected in an intelligent way. The algorithm is reported to be dynamically adaptive with increased detection rate while keeping the false alarm rate to the minimum.

Adebayo O et al [2008] have used two machine learning techniques namely Rough Set (LEM2) algorithm and k-nearest neighbour (kNN) algorithm for intrusion detection. However, poor detection rate of these algorithms on U2R and R2L attacks has been attributed to the few representations in the training dataset. But the attribute values in a training data set are completely different from the attribute values of the test dataset for these two attack types. Ozgur Depren et al (2005) have designed a model for both misuse and anomaly intrusion detection by employing SOM for detecting anomalies only with important but limited number of features. The model has been based only on normal behavioral patterns and any deviation from the normal is considered as an attack.

Zhi-song pan et al., [2003] have reported a misuse intrusion detection model based on a hybrid neural network and decision tree algorithm. They have discussed the advantages of different classification abilities of neural networks and the C4.5 algorithm for different attacks. While neural network algorithm is reported to have high performance to DOS and Probe attacks, the C4.5 algorithm has been found detect R2L and U2R attacks more accurately. In another study the same group [2005] has designed a hybrid approach combining expert system for misuse detection and back propagation neural network for anomaly detection. They have reportedly achieved 96.6 percent of detection rate for DOS and Probe with a false alarm rate of less than 0.04%. Expert system detects R2L and U2R more accurately than neural networks. They have concluded that the neural network could provide significant benefits to intrusion detection through data reduction, classification, clustering the unlabeled data and the process of identifying intruders.

Neural network algorithms have been employed for online pattern analysis (Da Deng and Nikola Kasabov, 2003). The system has been designed with null network and allows the network to grow with the help of connection strength and distance threshold. The random initialization of weight vector assignment in SOM has been modified in such a way that the original data space is assigned as weight vectors. It has been further reported that the network expands whenever the distance measured is more than the distance threshold. SOM has some limitations with real time applications such as fixed network architecture, dimensional reduction problem and lack of interpretability. In an attempt to overcome these limitations, Alahakoon D et al [2000] have presented an extended version of SOM with the advantage of discovering the knowledge in the network. The spread factor has been used as an essential parameter in controlling the growth of the network as it is independent of the dimensionality of data space. However, in this approach learning the map takes considerably long time as the learning rate has not been considered as a parameter.

In summary, none of the above said methods have solved completely the problem of fixed architecture and growth of the map in the SOM. Even though some of the methods have some advantages but it has its own draw backs such as undefined learning rate. Here we proposed combined approach which is modified from the simple SOM in terms of neighbourhood identification and finding the connection strength between the neighboring nodes along with k-means clustering analysis which refines the detection process. The map can be allowed to spread whenever needed with the help of threshold defined in trial and error method in the modified SOM and nodes of the resultant map are grouped by clustering algorithm using distance measure. As a result the modified SOM along with k-means has overcome the problem of fixed architecture and also wide spread nodes and also provides good detection rate.

# 3. PROPOSED WORK

In an attempt to further improve anomaly detection while reducing the false alarm rate, it has been decided in this work, by cascading modified SOM with k-means clustering algorithms. In modified SOM technique, the original data space was assigning as initial weight vectors encompassing more features instead of random initialization in which the number of output nodes are predefined. In the modified technique, it is expected that the number of output nodes are allowed to grow with the help of distance threshold. The wide spread nodes are further grouped into fixed number of clusters using k-means clustering analysis.

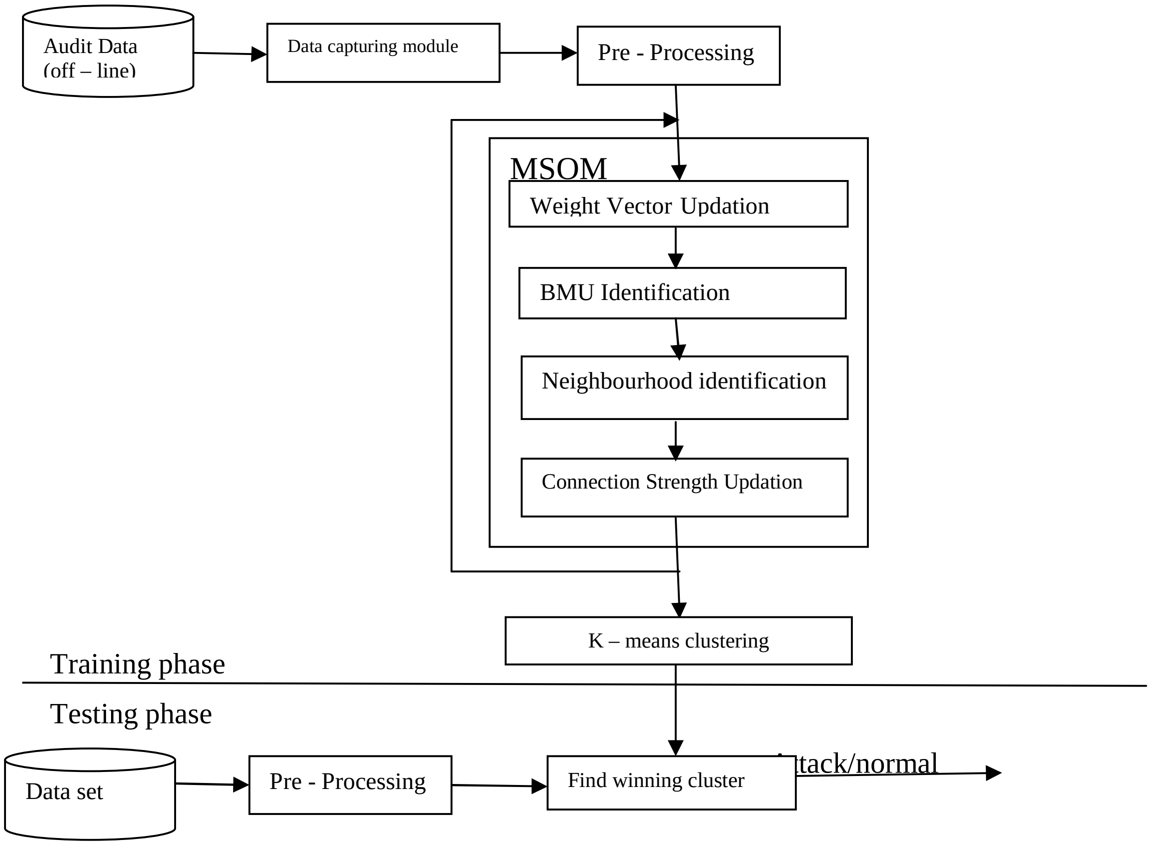

Figure 1. Proposed Architecture

In the combined frame work, the detection system is developed with three modules, namely, pre–processing phase, learning phase and testing phase. The learning phase comprised with two algorithms namely modified SOM and k-means are deployed to train the system, whose framework design is given in Figure 1.

An unsupervised audit data of network layer, which comprise of 41features are used to build the detection system. This data contains categorical, symbolic and continuous types of attributes. Creation of the system needs attribute values in numerical format for processing, and for getting the numeric value several pre-processing steps are needed to be performed.

In pre-processing module, the first step is to convert the data which is suitable for unsupervised learning by removing the labels from the dataset. In the second step, categorical attributes such as service type, protocol type, connection status flag were assigning numerical equivalent as given Table1, to perform the operations.

Table 1. Categorical Attributes vs Numeric Values

| Categorical attribute | Numeric value | Categorical attribute | Numeric value | Categorical attribute | Numeric value |
|---|---|---|---|---|---|
| tcp | 0 | S0 | 1 | S1 | 5 |
| udp | 1 | S3 | 2 | S2 | 6 |
| icmp | 2 | REJ | 3 | RSTR | 7 |
| sf | 0 | RSTO | 4 | SH | 8 |

In the third step normalization is carried out for the attributes using Min - Max technique (Han and Kamber, 2003) using the following formula:

$$V_{i\ (new)} = V_{i\ (old)} - V_{\ min} \ /\ V_{max} - V_{min}$$

Where $V_i$ is the new normalized value for i$^{th}$ record of the attribute, $V_{\ max}$ and $V_{\ min}$ is the maximum and minimum value of the attribute respectively.

The pre-processed data thus obtained was taken up by the learning module to start the system to learn. Initially, in the first phase of the learning module modified SOM, values for the distance threshold and learning rate were assigned with null network. As the data set enters and allowed the growth of the map, the best match unit and the distance measure were identified. The identified distance measure was compared with the distance threshold. Once a new node is created the activation is calculated using the equation (1) (Da Deng and Nikola Kasabov, 2003)and its connection strength with the other nodes are initialized to zero.

$$\mathbf{a_i= e^{-2\ ||x-w_i||\ \wedge\ 2\ /\ \varepsilon\wedge 2}} \qquad \textbf{eq.. (1)}$$

where $a_i$ is the activation value of the node i and ε is the distance threshold. The neighbourhood nodes were identified using the following neighbourhood function Ω(i) (Da Deng and Nikola Kasabov, 2003)as given in equation (2).

$$\mathbf{\Omega(i) = \{\ j\ /\ \ s(i, j) > 0\ \}} \text{ where } j = [1{:}\ n] \qquad \textbf{eq.. (2)}$$

Where n is the number of nodes and s(i, j) is the connection strength between node i and j. The connection strength between the neighbourhood nodes and the winner node was then updated using the formula [Da Deng and Nikola Kasabov, 2003] given in equation (3)

$$\mathbf{s_{new}(i, j) = \beta\ s_{\ old}(i\ , j) + (1 - \beta)\ a_i\, a_j} \qquad \textbf{eq.. (3)}$$

where β is the forgetting constant, $a_i$ and $a_j$ are the current activation values of node i and node j.

```
Modified SOM ALGORITHM:

Input: pre- processed dataset
Output: map with connection strength
The learning algorithm of modified SOM is given by following steps.

Step 1: A new input data vector x is taken.
Step 2: If there are no existing nodes or distance measure is greater than threshold then
        a. Create a new node and insert the input data as a weight vector.
        b. Find the activation value of the new node using equation (1) and its connection
           strength is initialized to 0.
Step 3: Find the BMU using the distance measure
Step 4: If distance measure of winner node is less than the distance threshold then
        (i) The activation value of the winner node is updated with equation (1),
        (ii)Neighbourhood nodes are identified using equation (2) ,
        (iii) Connection strength is updated and new weight vector is calculated using
              equation(3) and (4)
Else goto step 2.
Step 5:  Repeat until no more data are available.
```

Once neighborhood nodes are identified, then their weight vector values were updated along with winner node with the formula [Da Deng and Nikola Kasabov, 2003] given in equation (4)

$$\mathbf{W_i(t+1) = \gamma\ (a_i / \textstyle\sum_k a_k)\ (x- w_i\,(t))\ \ if\ i\ \text{€}\ \Omega(j)} \qquad \mathbf{(4)}$$

Where γ is the learning rate, t is the time and k is the number of nodes.

As the result number of nodes in the network is generated with neighbourhood function, connection strength and weight vector values. Secondly k-means clustering algorithm is deployed for clustering the nodes of network which are created in modified SOM with the help of distance measure. The process of forming clusters involves combining several variables into dissimilarity or distance measure whose values are then used to form groups (Lee et al., 2011). The set of nodes forms the k clusters using the k-means clustering algorithm as given below. The k-means clustering popularity was due to its fast convergence and simplicity.

```
K Means ALGORITHM
Input : nodes created in MSOM
Output: set of k clusters
Step1: initialize k value
Step 2: select k nodes as initial cluster seed values
Step 3: repeat
        For I = 1 to n
         Compute $|x_i – c_j|^2$ for all cluster seeds
          Assign $x_i$ to closest cluster $c_j$
          Re compute the cluster seed using mean function
        Until (no change in the cluster seed values)
```

In the test module new samples are taken and they are allowed to enter into pre-processing module for initial processing as described in the training phase. The processed data is given to the trained system to find the cluster which is created in the training module to find the winning cluster. According to the BMU cluster the data is considered as attack or normal.

## 4. EXPERIMENTS AND DISCUSSION

### 4.1 Dataset Description

Supervised anomaly detection dataset are taken from the standard bench mark dataset kddcup.data-10-percent –corrected in KDD cup99. The cup dataset contains more records of intrusion pattern using simulated environment to train the model. The network layer dataset which consist of 41 attribute which are considered as the features of that layer. From the training dataset three specific protocol records have been selected for learning the system. The system is trained with normal and attack dataset with tcp dump data as set I and icmp data as set II and udp data as set III. The dataset contains Normal, DOS , U2R & R2L and Probe as given in the Table 2.

Table 2 Dataset Description

| Data description | SET I | | SET II | | SET III | |
|---|---|---|---|---|---|---|
| | Normal | Attack | Normal | Attack | Normal | Attack |
| Total samples | 30800 | 2911 | 25900 | 2570 | 24320 | 2320 |
| Training samples | 24640 | 2328 | 20720 | 2056 | 19426 | 1856 |
| Testing samples | 6160 | 583 | 5180 | 514 | 4824 | 464 |

## 4.2 Discussion

The analysis has been carried out by combining modified self organizing map and k-means clustering analysis. By allowing the network to grow using modified self organizing map and the final network nodes are clustered in to k different cluster using the weight vector values of the each node. The analysis is performed by comparing the detection rate of the simple self organizing map, modified self organizing map individually along with proposed combined approach. The number of epochs and learning rate is selected by trial and error method. In modified SOM approach it has been evident from the result that it is possible to reach the stability in obtaining the weight vector with 100 epochs where as in simple SOM the number of epochs goes in thousand(Bose et. al.,2012). The reduction in the number of the epoch in the modified SOM saves considerable running time which is an important and desirable factor in intrusion detection.

Table 3 Comparison of SOM, MSOM and MSOM +k-means

| Testing samples | Correctly classified | | | Incorrectly classified | | |
|---|---|---|---|---|---|---|
| | MSOM | SOM | MSOM + k-means | MSOM | SOM | MSOM + k-means |
| Set I normal | 5922 | 5728 | 6036 | 238 | 432 | 124 |
| Set I attack | 560 | 542 | 571 | 23 | 41 | 12 |
| Set II normal | 5123 | 5102 | 5143 | 57 | 78 | 37 |
| Set II attack | 508 | 506 | 510 | 6 | 8 | 4 |
| Set III normal | 4799 | 4727 | 4809 | 25 | 97 | 15 |
| Set III attack | 461 | 454 | 462 | 3 | 10 | 2 |

The results are used for calculating the true positive, true negative, false positive and false negative values from which the detection rate (DR) and false positive rate (FPR) are determined. In Table 3 the performance of the test data set are analyzed and results are listed. The performance of the combined approach has been found to be better in terms of intrusion detection rate as well as the decrease in the false alarm rate. The Figure 2 gives graphical comparison of the performance of the modified SOM, simple SOM with combined approach.

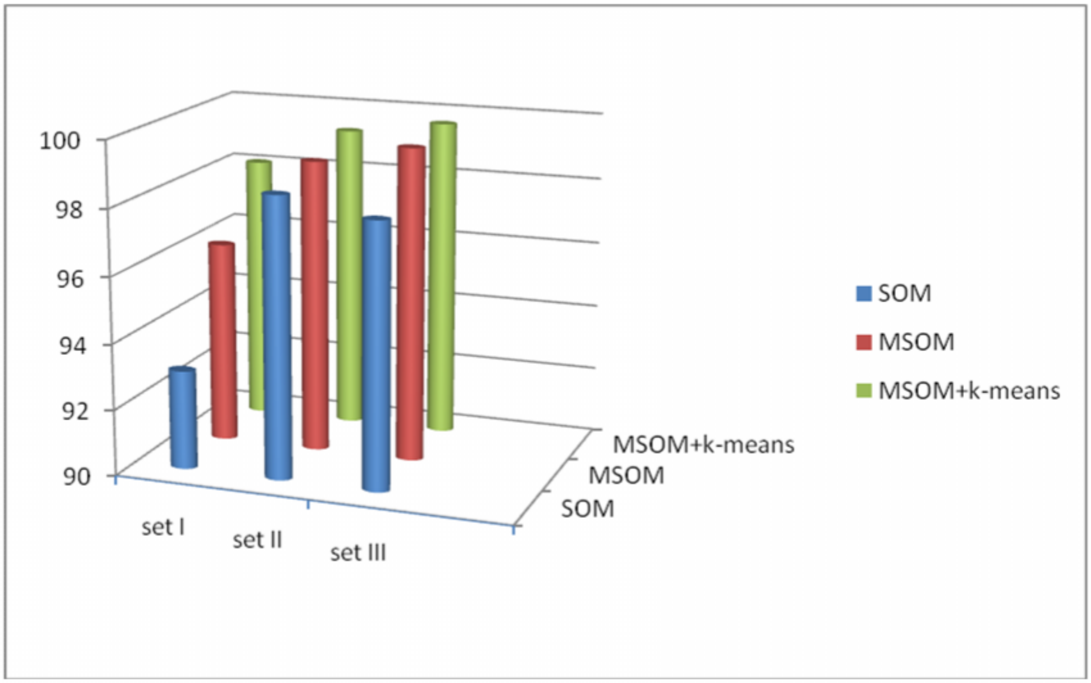

Figure 2. Comparison of Detection Rate

## 5. Conclusion

In this paper a combined approach using neural network and clustering algorithms for network anomaly detection is proposed. The modified SOM is used to create the network with the help of distances threshold, connection strength and neighbourhood functions and k-means clustering algorithms groups the nodes in the network with the help of similarity measures. The modified self organizing map has improved 2% higher detection rate compared to the existing SOM but when k-means is deployed it is further increased by 1.5%. It starts with null network and gradually evolves with original data space. The updation of neighbourhood function has been improved with the help of connection strength. The learning rate is found to plays the vital role by spreading the map as observed when the learning rate increases the number of output nodes decreases. In particular the proposed work is found to effective for detecting DOS attacks with 98.5% detection rate.

## References

[1] Adebayo O. Adetunmbi, Samuel O. Falaki, Olumide S. Adewale and Boniface K., (2008), "Network Intrusion Detection based on Rough Set and k-Nearest Neighbour", International Journal of Computing and ICT Research, Vol. 2(1), pp. 60 - 66.

[2] Alahakoon, D., Halgamuge, S. K., &Srinivasan, B., (2000), " Dynamic self-organizing maps with controlled growth for knowledge discovery", IEEE Transactions on Neural Networks, vol. 11(3), pp. 601–614.

[3] Bose. S, Aneetha . A.S, Revathi. S, (2012), "Dynamic network anomaly intrusion detection system using modified SOM", Proceedings of Second International Conference of Computer science and Engineering - 2012, New Delhi, pp. 27 - 34.

[4] Da Deng &Nikola Kasabov. N, (2003), "Online pattern analysis by evolving self-organizing maps", Elsevier, Journal of Neuro computing, vol. 51, pp. 87–103.

[5] Jiawei Han and Micheline Kamber, (2003), “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers.

[6] Juha Vesanto , Esa Alhoniemi,(2000), “Clustering of the Self-Organizing Map”, IEEE Transaction on Neural Networks, Vol. 11 (3), pp. 586 - 600.

[7] KDD cup 99 : Intrusion Detection Data set. < http:// kdd.jcs.uci.edu/databases/kddcup99/ kddcup.data_10_percent.gz>

[8] Ozgur Depren, Murat Toppallar, Emin Anarim, M.kemal Ciliz,( 2005), “ An Intelligent Intrusion Detection System (IDS) for anomaly and Misuse Detection in Computer Networks”, Elsevier, Expert System with Applications, vol. 29(4), pp. 713-722.

[9] Sandhya Pedabachigari, Ajith Abraham, Crina Grosan, Jhonson Thomas, (2007), “ Modeling Intrusion Detection System using Hybrid Intelligent Systems ”, Elsevier, Journal of Network and Computer Applications, vol. 30(1), pp. 114-132.

[10] Seungmin Lee, Gisung Kim, Sehum Kim, (2011), “Self – adaptive and dynamic clustering for online anomaly detection”, Elsevier, Expert System with Applications, Vol. 38(12), pp. 14891- 14898.

[11] Shekhar R. Gaddam, Vir V. Phoha, Kiran S. Balagani, ( 2007), “ K-means + ID3 : A novel method for supervised anomaly detection by cascading K- means clustering and ID3 decision tree learning methods”, IEEE Transactions on Knowledge and Data Engineering, vol. 19(3), pp. 345- 354.

[12] Villmann .T, Der . R, Hermann .M, Martinetz . M, (1997), “Topology preservation in self-organizing feature maps: Exact definition and measurement,” IEEE Transaction on Neural Networks, vol. 8 (2), pp. 256–266.

[13] Yasser Yasami, Saadat Pour Mozaffari, (2010), “A Novel Unsupervised Classification Approach for Network Anomaly Detection by K-Means Clustering and ID3 Decision Tree Learning Methods”, Springer, Journal of Supercomputing, vol. 53(1), pp. 231-245.

[14] Zhi-Song Pan, Song-Can Chen, Gen-Bao Hu, Dao-Qiang Zhang, (2010), “Hybrid Neural Network and C4.5 for Misuse Detection ” , Proceedings of the second International conference on Machine Learning and Cybernetics, November, pp. 2463 – 2467.

[15] ZhiSong PAN, Hong LIAN, GuYu HU, GuiQiang NI, (2005), “An Integrated Model of Intrusion Detection Based on Neural Network and Expert System”, Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI’05), pp. 2-3