Intrusion Detection System (IDS) Using Self Organising Map (SOM)

Prof.A.K.Gulve
GEC,Aurangabad
akgulve@yahoo.com

D.G.Vyawahare
ME,Final Year,CSE
G.E.C,Aurangabad
dgvyawahare@acm.org

ABSTRACT
The main objective developing Intrusion Detection system Using SOM is to help the network personals to achieve excellence in detecting network intrusion without fail. The basic idea behind using SOM for intrusion detection is its capability to even detect a small change in the data packets. The SOM will detect the anomaly behaviour even if it is trained on the very normal data set. This will make the designed system, a powerful tool for Intrusion Detection. The SOM is an unsupervised neural network algorithm that uses competitive learning [27]. Competitive learning means that as data is input to the SOM, there is a competition among the neurons or nodes of the map to determine which neurons will represent the input data. In the case of the SOM,the winning neuron is the neuron most similar to the input data,and it is affected by becoming more like the input data. In this way, neurons in the map become specialized to represent different sets of data in the input space.

1. Introduction
All data on the network travels in the form of packets, which is a basic data unit for network. The network layer is where the term packet is first time used. Common protocols at this layer are IP (Internet Protocol), ICMP (Internet Control Message Protocol), IGMP (Internet Group Management Protocol) and IPsec (Protocol Suite for securing IP). The transport layer protocols include TCP (Transmission Control Protocol), a Connection Oriented Protocol; UDP (User datagram protocol), a connection-less protocol; and SCTP (Stream Control Transmission Protocol), which has feature of both TCP and UDP.

In the past years, the networking revolution has finally come of age. More than ever before, we see that the Internet is changing computing as we know it. The possibilities and opportunities are limitless; unfortunately, so too are the risks and chances of malicious intrusions. It is very important that the security mechanisms of a system are designed so as to prevent unauthorized access to system resources and data. However, completely preventing breaches of security appear, at present, unrealistic. We can, however, try to detect these intrusion attempts so that action may be taken to repair the damage later. This field of research is called Intrusion Detection. Anderson, while introducing the concept of intrusion detection in 1980, defined an intrusion attempt or a threat to be the potential possibility of a deliberate unauthorized attempt to:

1. access information,
2. manipulate information, or
3. Render a system unreliable or unusable.

Capturing packets means collecting data being transmitted on network. We can capture these packets from the network using some open source libraries which support data capturing using packet sniffers. The intrusion to this network is detected using a technique called Self Organising Map (SOM). In the

bellow paper we see how the technique is used in the for intrusion detection.

2. LITERATURE SURVEY
2.1 Intrusion Detection System
Today large amount of data is passed over the network. Applications such as intrusion detection systems (IDS) become important to detect security threats. Intrusion detection monitors the events occurring in network or computer system. Intrusion Detection Systems implemented at the network-level or at the host-level. First works on individual packets to detect intrusion while other test the activity of individual host or computers. Data Mining based ISDs can be classified into two categories Misuse detection system and anomaly detection systems.

An Intrusion Detection system generally contains three functional components

1. An information source that provides a stream of event records
2. An analysis engine that detects intrusion and
3. A response component that generates outcome of analysis

Analysis engine takes information from data source and test it for sign of intrusion or attack. An analysis engine can use one or both of following approaches

Misuse Detection: Misuse detection system can detect only known attacks. Databases of well known pattern of attacks are compared to entries in this database.

Anomaly detection: Anomaly detection system search for something rare or unusual. It uses unlabeled data that helps the IDS in detecting new attacks.

2.2 Unsupervised Learning
Unsupervised Learning-Based Approaches: Supervised learning methods for intrusion detection can only detect known intrusions. Unsupervised learning methods can detect the intrusions that have not been previously learned. Examples of unsupervised learning for intrusion detection include K-means-based approaches and self-organizing feature map (SOM SOM-based approaches: Some

authors used the extract features that describe network behaviors from audit data, and they use the SOM to detect intrusions. Kayacik et al. propose a hierarchical SOM approach for intrusion detection. Specific attention is given to the hierarchical development of abstractions, which is sufficient to permit direct labeling of SOM nodes with connection type. In a hierarchical SOM for intrusion detection used the classification capability of the SOM on selected dimensions of the data set to detect anomalies.

Their results are among the best known for intrusion detection. Current approaches for intrusion detection have the following two problems.

a) Current approaches often suffer from relatively high false-alarm rates, whereas they have high detection rates. As most network behaviours are normal, resources are wasted on checking a large number of alarms that turn out to be false.

b) Their computational complexities are oppressively high. This limits the practical applications of these approaches.

Giovanni Vigna et al. (2003) developed a framework, called STAT, that supports the development of new intrusion detection functionality in a modular fashion. The STAT framework can be extended following a well-defined process to implement intrusion detection systems tailored to specific environments, platforms, and event streams.. The resulting intrusion detection systems represent a software family whose members share common attack modeling features and the ability to reconfigure their behavior dynamically.

They evaluated the performance impact of the framework based approach by comparing the performance of the original, *ad hoc* version of NetSTAT to the one developed by extending the STAT framework. The two systems were ran on a file containing two days worth of network data from the 1999 MIT Lincoln Laboratory evaluation. The total CPU time was collected for both sensors during multiple runs. The average processing time was 3,220 seconds for the original NetSTAT and 2,862 seconds for the framework based sensor. The speedup of 13.8% is attributed to careful optimization of the framework source code. (Data Mining Based)

Stefano Zanero et al. (2004) proposed a novel architecture which implements a network-based anomaly detection system using unsupervised learning algorithms. They described how the pattern recognition features of a Self Organizing Map algorithm can be used for Intrusion Detection purposes on the payload of TCP network packets. They used a two-tier architecture, which allows us to retain at least part of the information related to the payload content. Their final goal was to detect intrusions, separate packets with anomalous or malformed payload from normal packets.

The prototype was ran over various days of the 1999 DARPA dataset. A 66.7% detection rate with as few as 0.03% false positives was obtained. The detection rate was maximum upto 88.9% for threshold 0.09% with a false positive rate 0.095%.

Liberios Vokorokos (2006) , presented intrusion detections systems and design architecture of intrusion detection based on neural network self organizing map. Result of the designed architecture is simulation in real conditions [3]. The goal of the proposed architecture was to investigate effectiveness of application of a neural network at modelling user behavioural patterns so that they can distinguish between normal and abnormal behaviour. Expected network reply was the value close to-for user, which behaviour not diverting from normal behaviour. If the output value of network becomes above specified threshold value, alarm was raised.

The results were obtained on the department server KPI Technical University of Košice. Neural network SOM in the IDS systems. Collecting of essential information from single controlled points lasts 2 days. Next the neural network SOM was created and trained, which serves as the core of the IDS system. The results shown that input vectors classification, which represents behaviour and its mapping to particular neurons, form single possible user behaviour states. Formed states were as intrusion – Intrusion, possible intrusion – Intrusion?,

H. Günes Kayacık et al.(2006) focused on developing behavioral models of known attacks to help security experts to identify the similarities between attacks. A Self Organizing Feature Map (SOM) was employed to model the relationship between known attacks and UMatrix representation was used to create a two dimensional topological map of known attacks. The approach was evaluated on KDD'99 data set. Results showed that attacks with similar behaviour patterns are placed together on the map.

Considering the dataset needs to be balanced to eliminate any bias towards majority classes, they trained a Self-Organizing Map on the balanced training data and employed the labels (i.e. attack types) from the same dataset to assign labels to neurons. The concept of a best matching node was used to facilitate the labelling of the map.

Results on the test data indicate that known attacks are identified with relatively high identification accuracy although SOM employs unsupervised learning.

By using KDD 10 % dataset accuracy of attacks like perl, smurf, back, nmap found to be 100%,99.99%,88.24% and 48.48% respectively and that with corrected dataset accuracy of attacks like perl reduced to 50%, whereas back & nmap increased to 100%.

Zhenwei YU et al. (2008), They presented an automatically tuning intrusion detection system, which controls the number of alarms output to the system operator and tunes the detection model on the fly according to feedback provided by the system operator when false predictions are identified. The system was evaluated using the KDDCup'99 intrusion detection dataset.

They proposed an adaptive and automatically tuning intrusion detection system, ADAT.

The results shown demonstrated that the ADAT model tuner improved the overall classification accuracy while decreasing total misclassification cost. Compared to the multi classifier SLIPPER-based IDS without the tuning feature, ADAT reduced total misclassification cost (52294 as compared to 70177 of MC-Slipper) by 25.5%, while increasing overall accuracy by 1.78%. Compared to the automatically tuning IDS with delayed tuning, ADAT reduced TMC by 6.76%.

Stefano Zanero (2008) , presented a tool for network anomaly detection and network intelligence which was named as ULISSE. It uses a two tier architecture with unsupervised learning algorithms to perform network intrusion and anomaly detection. ULISSE uses a combination of clustering of packet payloads and correlation of anomalies in the packet stream.

In order to evaluate the architecture in a repeatable manner, the prototype was ran over various days of track drawn from the 4th week of the 1999 DARPA dataset [14]. They also added various attacks against the Apache web server and against the Samba service generated through the Metasploit framework (www.metasploit.org).

It was concluded that their architecture can reach the same detection rate of 66.7% (PAYL [15]) with a false positive rate below 0.03%, thus an order of magnitude better than PAYL, or on the other hand reach a 88.9% detection rate with no more than a 1% rate of false positives.

V. K. Pachghare et al.(2009) developed their own packet sniffer. Apart from capturing live packetsthey also used a standard DARPA dataset, for training purpose [17]. The dataset contain both packets with intrusion and without intrusion. The accepted window length was 20 for the application. Since the data were collected in every 20 seconds an input vector corresponds to time interval of 400 seconds.

For training purpose they constructed a 30x30 Self Organizing Map in order to perform clustering. The data that was used for it was DARPA dataset [17]. Batch training algorithm with training length 100 and starting radius 15 was used. Self organizing map was found largely successful in classifying the IP packets. After the data collection, vector extraction and training of the Self Organizing Maps, the packets were passed through the SOM. The result was shown in form of patterns.

They concluded that, the actual experiments show that even a simple map, when trained on normal data, can detect the anomalous features of both buffer overflow intrusions exposed to it. This approach found particularly powerful because the self organizing map never needs to be told what intrusive behavior looks like [18].

Mansour M. Alsulaiman et al. (2009) they built an Intrusion Detection System using a well known unsupervised neural network, namely Kohonen maps. They proposed two enhancements that were able to solve one of the shortcomings of the available solutions, namely high value of false positive rate. The method called as Performance-Based Ranking Method [21] was used. It works by deleting an input from the dataset and comparing the result before and after the deletion. They used the KDD data set which is available in [20]. To make the data in the right format, as an input to their system, they changed some of its feature formats, because neural network accept only numeric data. They changed 3 features, namely the protocol, flag and service to numeric data.

After this they tried to find ways to improve the results by proposing and investigating several enhancements to HSOM. HSOM was a powerful improvement to SOM, so they used it and got some good results. Thus they found ways to improve it. One enhancement was to complement it with PBRM and good results were obtained. Another enhancement was to add more layers. They showed that by good analysis and selecting the best layer to compliment a combination better result can be obtained.

The two enhancements were presented :

A. HSOM with PBRM: They applied the unresolved patterns of Net3 to a trained PBRM network; The PBMR classified the unresolved patterns into normal or attack with a recognition rate of 99% and a false positive rate of 2.25%.

B. New combination: They created a new combination by adding a new layer. The new layer can be a layer from another combination. They postulate that, if this layer is chosen to be the layer responsible for resolving the largest number of neurons, then that can help the other combination.

The proposed enhancement did improve the result. HSOM with PBRM improved the recognition rate from 94.93% to 99%, and gave an acceptable false positive rate, namely 2.25%.

In this work it was shown that SOM is an excellent choice to build IDS.

2.2.1 SOM

Self Organising Map (SOM) is a type of artificial neural network which is trained using unsupervised learning technique. SOM is generally used to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map**[21].** The SOM is consisting of nodes called as neurons. Each neuron is having a weight age assigned to it[21]. The SOM is generally a procedure by which a multi dimension data input data is mapped to two dimension data. Each node in the map is traversed to find the similarity between the input vector and the map's node's weight vector using a Euclidean distance formula. The Euclidean formula is

$$\|x - m_c\| = \min \{\|x - m_i\|\}$$

The node producing the smallest distance is tracked. This node is called as Best Matching Unit (BMU). The nodes neighbouring the BMU are updated by tacking them closer to the input vector[21].

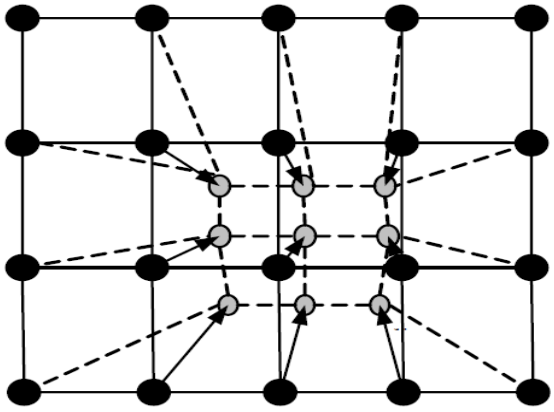

Fig. 1 General SOM Topology

3. IMPLEMENTATION OF SOM TO IDS

The SOM can be applied to the Intrusion Detection System to get the intrusion in the system. The steps involved in the application of SOM to Intrusion detection System is given as below-

3.1 Packet Sniffing

Packet sniffing means capturing of the data packets sent over the network. Every time a card receives an Ethernet frame, it checks

if its destination MAC address matches its own[22]. If it does it generates an interrupt request. The routine that handles this interrupts is the network card drivers; it copies the data from card buffer to kernel space, then checks the *ethertype* field of the Ethernet header to determine the type of the packet, and passes it to the appropriate handler in the protocol stack[22]. The data is passed up the layer until it reaches the user-space application, which consumes it.

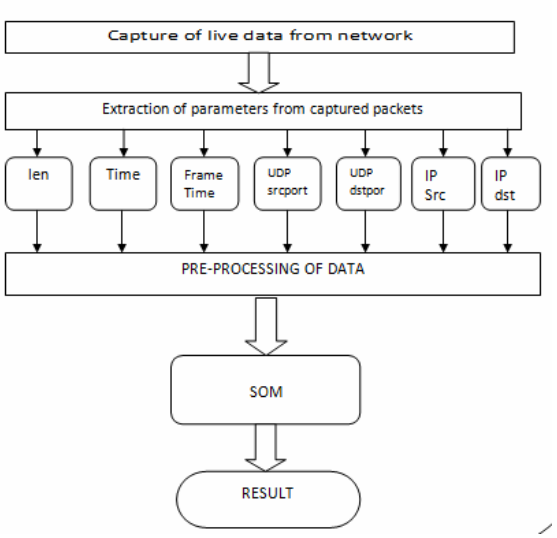

Fig. 2 SOM Application to IDS

For packet sniffing, we are using *libpcap*, and open source library. Libpcap is a platform-independent open source library to capture packets. (Windows Version of libpcap is winpcap). Famous sniffers like tcpdump and wireshark make use of this library.

3.2 Extraction of Parameters

The captured data packets having some parameters that can be used for the SOM[26]. Of the many parameters possessed by the packets, we are taking only seven of them to obtain the multi dimension data for SOM processing. The properties or parameters used in this capture are

- *Len:* This is length of the data packet captured.
- *Time:* This is time stamp of the data packet captured. It typically has the values like creation of the data packet and etc.
- *Frame T*ime: It is the time constraint of the individual frame taken in the training.
- Udp srcport: It is the UDP packet source port given in the captured packet.
- UDP dstport: It is a UDP packet destination port mentioned in the captured data.
- IP srcport: It is a source port value given in the IP packet.
- IP dstport : It is the IP destination port given in the captured IP packet.

This extracted data is then saved in the comma separated values type file[26].

4. PRE-PROCESSING

Pre processing of the data is done on the extracted features of captured packets. Data pre processing includes analysis of the data and assignment of value to the null fields. Analysis includes checking of the csv file for appropriate number of fields in the file. In this case we are considering the seven features of the captured data packets[21].

If any null value is found for extracted data then in the pre-processing phase, it will assign a zero value to that field. This will make the data ready to use in the next step.

4.1 Som Application To The Data

After pre-processing of the data, SOM is applied on the data. By the definition of SOM, it will convert the multi dimensional data in to two dimensional data. In this case all the data is to be in the numerical format. On this extracted and pre processed data, SOM is applied to train the network. This include finding the distance between the nodes in the map using the Euclidean formula and then finding the BMU by pulling these nodes closer[23].

4.2 Algorithm

1. Randomize the map's nodes' weight vectors
2. Grab an input vector
3. Traverse each node in the map
    1. Use Euclidean distance formula to find similarity between the input vector and the map's node's weight vector
    2. Track the node that produces the smallest distance (this node is the best matching unit, BMU)
4. Update the nodes in the neighbourhood of BMU by pulling them closer to the input vector

    $$\mathbf{Wv}(t+1) = \mathbf{Wv}(t) + \Theta(t)\alpha(t)(\mathbf{D}(t) - \mathbf{Wv}(t))$$

5. Increase t and repeat from 2 while $t < \lambda$

5. EXPERIMENTAL RESULTS

We have captured the real time data from the network for this simulation. We have captured various data packets from college network using packet capture software programmed in java. This software is using a Winpcap library for packet detection and captured. We have taken various data set. Each of containing 600 to 10000 packets. We have also taken DARPA data set for more accuracy in the experimental result.

This will contains clean as well as intruded packets. These packets are then supplied to the designed system. After processing on the system, results are obtained from the system in the form of graphs. Results are shown below.

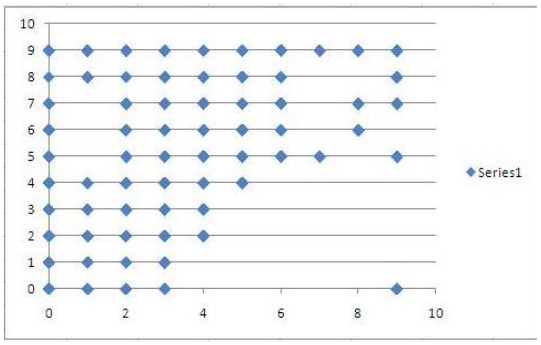

Fig. 3 Data without intrusion

In the above shown fig., a data packets of is given to the system. This packet is without intrusion. We can see the obtained graph for these data packets. The scatter graphs show various values obtained after application of SOM over extracted data from these captured packets. In this graph various values are compared among them.

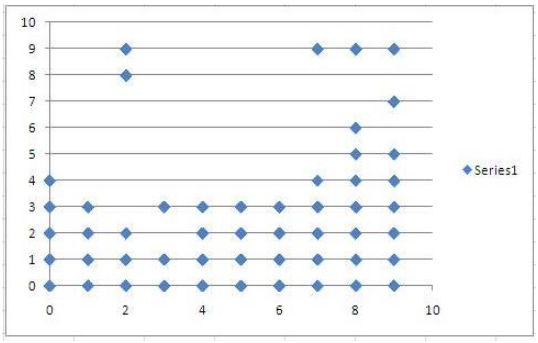

Fig. 4 Graph for Intruded data

Fig. 4 shows the same sample result but with intrusion. We can easily compare from fig. 3 & 4 that which graph is having intrusion. We can see that in fig. 4 some points in the graphs are missing and some are added to the graph newly. We have also done this operation on several sets of captured data and found the desired result.

For standardising our system we have also done the same operation on the DARPA data. We have taken the DARPA data and it is then supplied to the designed system. The graphical result for the DRAPA data is given below.

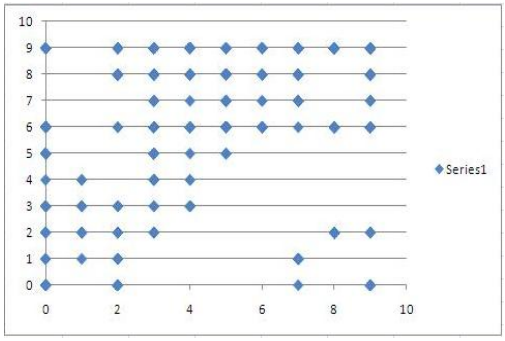

Fig. 5 DARPA DATA

In the above result we can see the pattern of the graph. This result is of 2000 data packets taken from the DARPA data set.

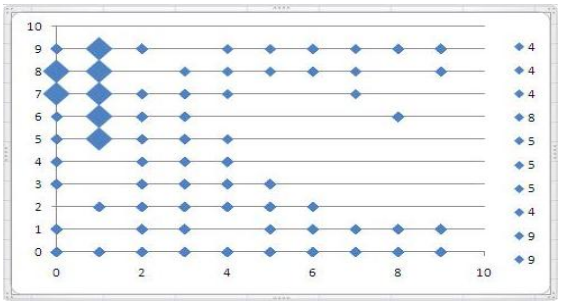

Fig. 6 DARPA DATA result

Fig.6 shows another result of the test data from the DARPA dataset. By comparing two figures we can found that in the fig 6, points marked bold are new data points. These points are not present earlier in the fig. We can also see the changed graph state form the two figures. These clear results will make the SOM, a better choice for use in the Intrusion Detection system.

6. CONCLUSION

Now a day's computer and network security is the crucial factor in the field of computer. Many intrusions to the network are done by hackers. This will make network, unsafe place for our data.

The goal of developing this system is to help the network personals to achieve excellence in detecting network intrusion without fail. The basic idea behind using SOM for intrusion detection is it capability to even detect a small change in the data packets. The SOM will detect the anomaly behaviour even if it is trained on the very normal data set. This will make the designed system, a powerful tool for Intrusion Detection.

7. FUTURE SCOPE

In this designed system, we are only considering seven parameters of a data packet. As we know a data packet has 40+ parameters. By considering, more than 10 or 12 parameters, we can apply SOM on them to get more appropriate result. This increased data collection will definitely increase the precision value of the result.

Due to use of more parameters from data packet, we are getting more nodes on the map. After applying SOM method on the extracted data, we can even get more realistic data.

8. REFERENCES

[1] A.K.Gulve and D.G.Vyawahare, Survey On Intrusion Detection System, *International Journal of Computer Science and Applications, 4(1)*, April/ May 2011, ISSN: 0974-1003 pages 7-13.

[2] Giovanni Vigna Fredrik Valeur Richard A. Kemmerer, Designing and Implementing a Family of Intrusion Detection Systems, *ESEC/FSE'03,* September 1–5, 2003, Helsinki, Finland. ACM 1-58113-743-5/03/0009

[3] Stefano Zanero,Sergio M. Savaresi, Unsupervised learning techniques for an intrusion detection system, *SAC'04* March 1417, Nicosia, Cyprus, ACM 1581138121/03/04.

[4] Liberios VOKOROKOS, Anton BALÁŽ, Martin CHOVANEC, Intrusion Detection System Using Self Organising Map,Acta Electrotechnica et Informatica No. 1, Vol. 6, 2006 ,pp.1-6

[5] H. Günes Kayacık, A. Nur Zincir-Heywood, Using Self-Organizing Maps to Build an Attack Map for Forensic Analysis, PST 2006, Oct 30-Nov 1, 2006, Markham, Ontario, Canada, ACM 1-59593-604-1/06/00010.

[6] Zhenwei Yu, Jeffrey J. P. Tsai, An Automatically Tuning Intrusion Detection System, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 37, NO. 2, APRIL 2007,pp. 373-384.

[7] Stefano Zanero, ULISSE, a Network Intrusion Detection System, CSIIRW '08 May 12-14, Oak Ridge, Tennessee, USA ACM 978-1-60558-098-2

[8] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba,and K. Das. Analysis and results of the 1999 DARPAo_-line intrusion detection evaluation. In Proceedingsof the Third International Workshop on Recent Advances in Intrusion Detection, pages 162{182,London, UK, 2000. Springer-Verlag.

[9] K. Wang and S. J. Stolfo. Anomalous payload-basednetwork intrusion detection. In RAID Symposium,September 2004.

[10] V. K. Pachghare, Parag Kulkarn, Deven M. Nikam, Intrusion Detection System Using Self Organizing Maps, 978-1-4244-4711-4/09/2009 IEEE.

[11] McHugh, J.: Testing intrusion detection systems: a critique ofthe1998 and 1999 DARPA intrusion detection system evaluationsas performed by lincoln laboratory. ACM Trans. on Informationand System Security 3 (2000) 262-294.

[12] Lane, T., and Brodley, C. E. 1999. Temporal sequence learning and data reduction for anomaly detection. ACM Transactions on Information and System Security 2(3):295-331.

[13] Mansour M. Alsulaiman, Aasem N. Alyahya, Raed A. Alkharboush, Nasser S. Alghafis, Intrusion Detection System using Self-Organizing Maps, 2009 Third International Conference on Network and System Security, 978-0-7695-3838-9/09 ,DOI 10.1109/NSS.2009.62

[14] http://www.securityfocus.com/infocus/1520 - An introduction to IDS,(last checked 15/July/2009)

[15] Srinivas M., Andrew H., Features Selection for Intrusion detection using Neural Networks and Support Vector Machines,Transportation Research Board, winter 2003.

[16] Stefan Axelsson, Combining a Bayesian Classifier with Visualisation: Understanding the IDS, *VizSEC/DMSEC'04,* October 29, 2004, Washington, DC, USA., ACM 1581139748/04/0010

[17] E. R. Tufte. The Visual Display of Quantitative Information. Graphics Press, second edition, May 2001. ISBN 0–96–139214–2.

[18] Iftikhar Ahmad, Azween B Abdullah, Abdullah S Alghamdi, Application of Artificial Neural Network in Detection of DOS Attacks, *SIN'09,* October 6–10, 2009, North Cyprus, Turkey. ACM 978-1-60558-412-6/09/10.

[19] Iftikhar Ahmad, M.A Ansari, Sajjad Mohsin. "Performance Comparison between Backpropagation Algorithms Applied to Intrusion Detection in Computer Network Systems" in the

[20] Book RECENT ADVANCES in SYSTEMS, COMMUNICATIONS & COMPUTERS, Included in ISI/SCIWeb of Science and Web of Knowledge & as ACM guide,2008, pp 47-52.

[21] Iftikhar Ahmad, Sami Ullah Swati, Sajjad Mohsin. "Intrusion Detection Mechanism by Resilient Back Propagation (RPROP)" EUROPEAN JOURNAL OF SCIENTIFICRESEARCH, Volume 17, No. 4 August 2007,pp 523-530.

[22] Antonis Papadogiannakis, Michalis Polychronakis, Evangelos P. Markatos, Improving the Accuracy of Network Intrusion Detection Systems Under Load Using Selective Packet Discarding, *EUROSEC* '10, Paris, France , 2010 ACM 978-1-4503-0059-9/10/04.

[23] Liberios Vokorokos, Anton Balaz, Martin Choveneck: Intrusion Detection System using self organizing map.

[24] Lane Thames: The use of Self Organizing Maps for intrusion detection.

[25] Brandon Craig Rhodes, James, A. Mahaffey, James D. Cannady : Multiple Self Organizing Maps for Intrusion Detection.

[26] Stefano Zanero, Sergio M. Savarasi : Unsupervised Learning Techniques For an Intrusion detection system, 2004 ACM symposium on Applied Computing.

[27] Peter Lichodzijewski, A.Nur Zincir-Heywood, Malcolm I. Heywood: Dynamic intrusion detection using self-organizing maps

[28] Samuel Kaski Data exploration using self organizing maps. Acta Polytechnica Scandinavica Mathematics Computing and Management in Engineering Series No .82 ,March 1997