DYNAMIC NETWORK ANOMALY INTRUSION DETECTION USING MODIFIED SOM

Aneetha.A.S., Revathi.S., Bose.S

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

ABSTRACT

Detection of unexpected and emerging new threats has become a necessity for secured internet communication with absolute data confidentiality, integrity and availability. Design and development of such a detection system shall not only be new, accurate and fast but also effective in a dynamic environment encompassing the surrounding network. In this paper, an algorithm is proposed for anomaly detection through modifying the Self – Organizing Map (SOM), by including new neighbourhood updating rules and learning rate dynamically in order to overcome the fixed architecture and random weight vector assignment. The algorithm initially starts with null network and grows with the original data space as initial weight vectors. New nodes are created using distance threshold parameter and their neighbourhood is identified using connection strength. Employing learning rule, the weight vector updation is carried out for neighbourhood nodes. Performance of the new algorithm is evaluated for using standard bench mark dataset. The result is compared with other neural network methods, shows 98% detection rate and 2% false alarm rate.

KEY WORDS

Anomaly Detection, Learning Rate, Weight Vector, Neighbourhood Function

1. INTRODUCTION

The basic function of any intrusion detection system is to detect inappropriate, inaccurate and anomalous activity in a system. Attacks may be in any form such as denial of service, remote to user, user to root 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 and anomaly detection are very often employed (Gaddam et al., 2007). While the former is capable of identifying known attack patterns, the latter depends only on the normal behavior and any deviation from the normal behavior is classified as an anomaly.

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. With appropriate

David C. Wyld, et al. (Eds): CCSEA, SEA, CLOUD, DKMP, CS & IT 05, pp. 27–34, 2012.
 DOI : 10.5121/csit.2012.2204

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. An attempt has been made in this work to rectify this drawback by proposing an algorithm that allows the network to grow, with a distance threshold, and also by using the connection strength to identify the neighbourhood nodes.

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.

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 the proposed algorithm, which is modified from the simple SOM in terms of neighbourhood identification and finding the connection strength between the neighboring nodes. The map can be allowed to spread whenever needed with the help of threshold defined in trial and error method. The learning rate and activation value of the nodes are used to calculate the connection strength. As a result the modified SOM has overcome the problem of fixed architecture.

3. PROPOSED WORK

In an attempt to further improve anomaly detection while reducing the false alarm rate, it has been decided in this work, to modify the SOM technique by assigning the original data space 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.

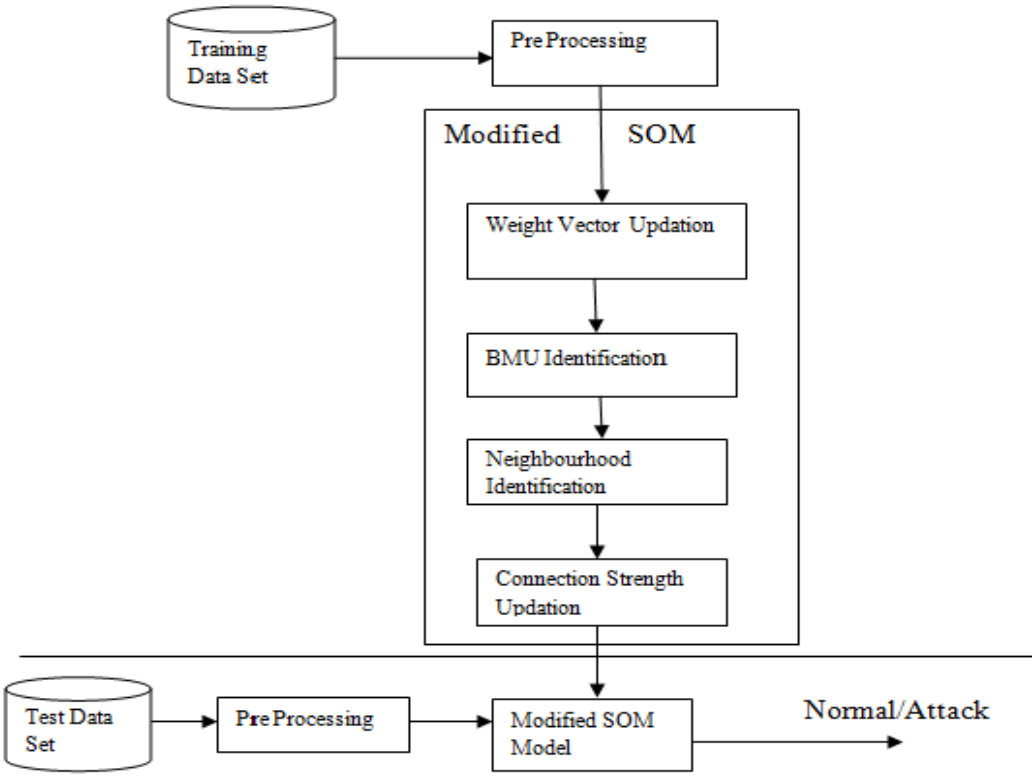

Figure1: Proposed Framework for Anomaly Detection

In the modified SOM frame work, the detection system is developed with three modules, namely, pre–processing, training and testing as shown in Figure 1. Audit data from a network layer with 41features are used to build the map. This data contain categorical, symbolic and continuous types of values. Creation of the map needs data in numerical format and for getting the numeric value several pre-processing steps are required to be performed.

In pre-processing module, the first step was to convert the data to a form suitable for unsupervised learning and this was done by removing the labels from the dataset. In the second step, categorical and symbolic data were transformed by assigning numerical equivalent to perform the operations. Normalization was then carried out for the selected features using Min Max technique (Han and Kamber, 2003) using the following formula:

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

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

The pre-processed data thus obtained was taken up by the training module to start the system to learn. Initially, 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.

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

where a_i is the activation value of the node 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).

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

Where n is the node of nodes and s(i, j) is the connection strength between node i and node 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)

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

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

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)

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

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

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 final prototype which is created in the training module to find the BMU. According to the BMU node 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 tcpdump 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 1.

Table 1 Dataset Description

Data set	Normal	DOS	U2R &R2L	Probe
Set I	70%	15%	10%	5%
Set II	75%	10%	10%	5%
Set III	80%	10%	5%	5%

4.2 Discussion

The analysis has been carried out by allowing the map to grow using modified self organizing map and the results are compared with the simple self organizing map. The number of epochs is selected by trial and error method. In this 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. 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 2 Comparison of SOM and MSOM

Test	Detection Rate(%)		False Positive Rate(%)	
	Proposed	Existing	Proposed	Existing
Set I	96.14	93	3.86	7
Set II	98.9	98.5	1.1	1.5
Set III	99.5	98	0.5	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. The results are listed in the Table 2. The performance of the modified SOM algorithm 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 and Figure 3 gives graphical comparison of the performance of the modified SOM with the simple SOM.

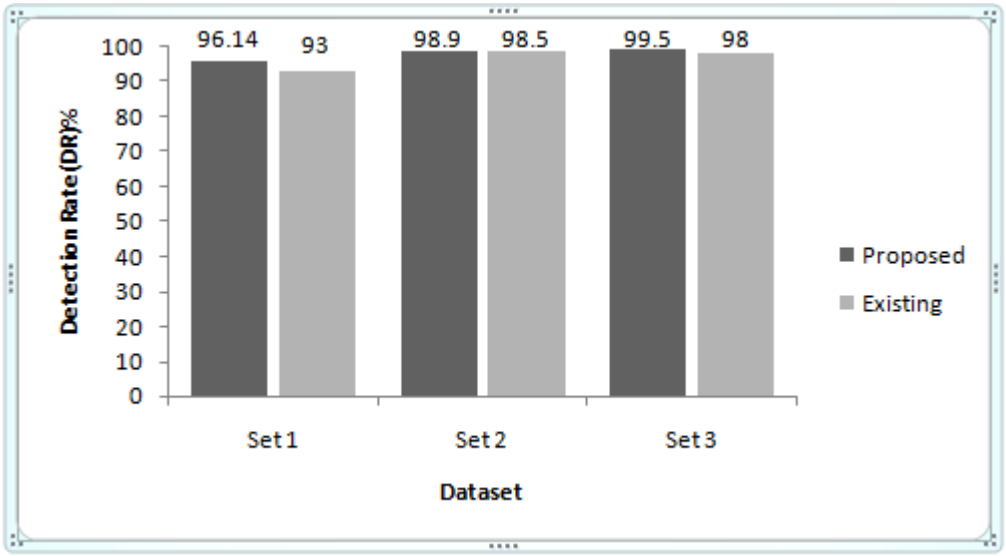

Figure 2 Performance of Algorithm for Detection Rate

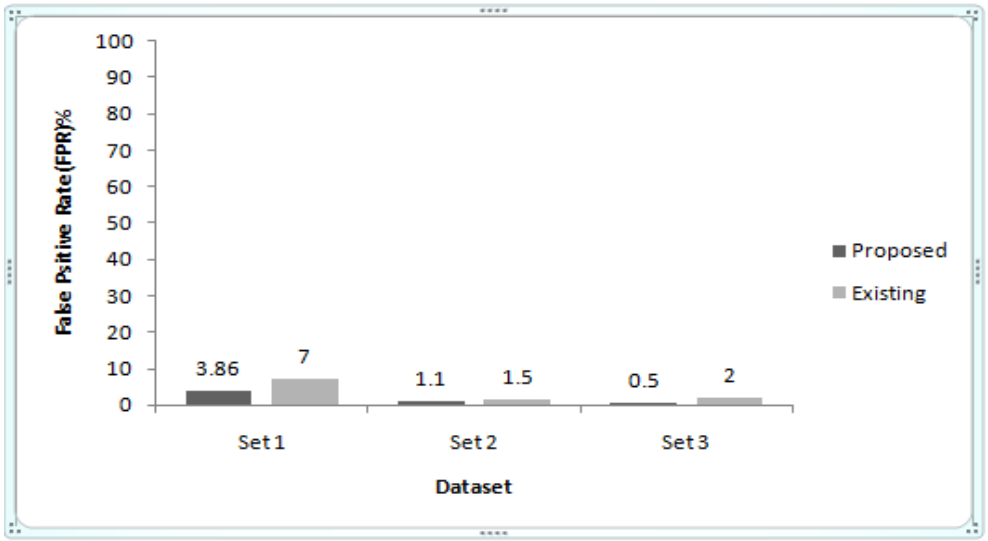

Figure 3 Performance of Algorithm for False Positive Rate

5. CONCLUSION

In this paper a dynamic framework for network anomaly detection is proposed. This modified SOM has improved 2% higher detection rate compared to the existing SOM. 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] Da Deng&Nikola Kasabov. N, (2003), "Online pattern analysis by evolving self-organizing maps", Elsevier, Journal of Neuro computing, vol. 51, pp. 87–103.

[4] Jiawei Han and Micheline Kamber, (2003), "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] Shekhar R. Gaddam, Vir V. Phoha, Kiran S. Balagani, March 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.

[11] 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.

[12] 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.

[13] 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.