IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

IMPROVING USER-TO-ROOT AND REMOTE-TO-LOCAL ATTACKS USING GROWING HIERARCHICAL SELF ORGANIZING MAP

Kruti Choksi*, Prof. Bhavin Shah, Asst. Prof. Ompriya Kale

[1]Student of Final Year M.E.(C.E.), L.J.Institte of Engineering and Technology, Ahmedabad.
[2]M.C.A. Programme, L.J. Institute of Management Studies, Ahmedabad.
[3]Department of Computer Engineering, L.J. Institute of Engineering & Technology, Ahmedabad, India.
kruti_492@yahoo.co.in

ABSTRACT

Intrusion Detection System (IDS) protects a system by detecting "known" as well as "unknown" attacks and generates the alert for suspicious activities in the traffic. There are various approaches for IDS, but our survey was focused on IDS using Self Organizing Map (SOM). Our survey shows that the existing IDS based on Self Organizing Map (SOM) have more computational time and poor detection rate for User-to-Root (U2R) and Remote-to-Local (R2L) attacks. So, our objective is to improve the detection rate of U2R and R2L attacks along with low computational time. From our survey we found that, Growing Hierarchical Self Organizing Map (GHSOM) is efficient due to its low computation time compared to traditional SOM model. To achieve our objective, our model uses GHSOM algorithm along with proper features selection to improve the performance of U2R and R2L attacks. Our empirical result indicates that, there is nearby 75% increase in the detection rate of U2R and R2L attacks by using GHSOM approach compare to SOM approach.

KEYWORDS: Intrusion Detection System (IDS), Self Organizing Map (SOM), User-to-Root (U2R), Remote-to-Local (R2L), Growing Hierarchical Self Organizing Map (GHSOM).

INTRODUCTION

An Intrusion Detection System (IDS) is a standard security component, which can detect security breaches and notifies to the administrator about the computer attacks, so that proper action can be performed [1] [3]. Various approaches are available for IDS; like Data Mining, Neural Network, and Genetic and so on [1] [2]. Amongst them Neural Network Approach proved efficient for IDS [1] [25], as it has high accuracy for detecting both "known" as well as "unknown" attacks.

The Self-Organizing Map (SOM) is a neural network model for analyzing and visualizing high dimensional data to one or two dimension data [15]. SOM proved efficient for IDS but faces some challenges. The neuron grid of SOM is predetermined, so sometimes it happen that many neurons do not take part in detection process and unnecessarily increases the computational time [7][10][11][12]. Another drawback of SOM is that, there is no standard technique to select the neuron grid size; it is done through trial and error basis which is also a time consuming factor. To overcome these challenges of SOM, Growing Hierarchical Self Organizing Map (GHSOM) is introduced in which the neuron grid grows dynamically as the data is given as an input to the grid [7] [10] [11] [12]. This dynamic nature of GHSOM neuron grid will not consist of any waste neuron; so computationally it will require less time compared to SOM.

Our survey [24] showed that the root cause of poor detection rate of User-to-Root (U2R) and Remote-to-Local (R2L) attacks are: (1) loss of integrity due to normalization of data [5], (2) similarity in records of normal and R2L attacks [6] [15] and (3) the records of U2R and R2L attacks are much fewer in comparison to Normal, DoS and Probe attacks records in KDD cup 1999 dataset which can affect the training process. Also our survey suggests that, feature selection is advantageous to improve the detection rate of IDS.

In this paper, our proposed work is focused on improving the detection rate of U2R and R2L attacks using GHSOM approach.

RELATED WORK

During our literature survey in [24], it was identified that SOM and its models are widely used for the IDS with different objectives.

The SOM approach was used to detect the anomalies by Pachghare [4]. While for improving the detection rate, Wang [5]introduced a co-relation parameter in learning rules of SOM. Gunes Kayacik [12] and Alsulaiman [3] used HSOM to build IDS for higher performance with higher anomalies detection and reduced false positive rate.

Ortiz [11], Palomo [10] and Mansour[7] approach was to improve the performance of IDS using GHSOM .Ortiz [11] built IDS to improve the detection rate using GHSOM along with probability labelling method. Palomo [10] used GHSOM to deal with the symbolic data and improving the detection rate. Mansour [7] used GHSOM to reduce the false alarms. But Ippoliti [6] [9] and Salem [8] along with the performance improvement of IDS also focus on the stability of GHSOM. But the problem of poor detection rate of U2R and R2L attacks existed in the majority of the IDS based on SOM models.

The Wilson [13] worked to improve the detection rate to U2R and R2L attacks, which is quite similar to our objective. Wilson[13] have achieved 13.8% detection rate for U2R attacks but fail to detect the R2L attacks using SOM algorithm with hotspot vector and vector pruning technique. In our proposed work we are also improving the detection rate of U2R and R2L attacks by using GHSOM approach along with relevant features selection.

PROPOSED WORK

Our proposed model to improve the performance of U2R and R2L attack is shown in figure 1. Our model uses benchmark dataset for IDS i.e. KDD cup 1999 dataset [3] [10] [13]. The KDD cup 1999 dataset need to be pre-process before usage as it contains symbolic data which are not acceptable by neural network.

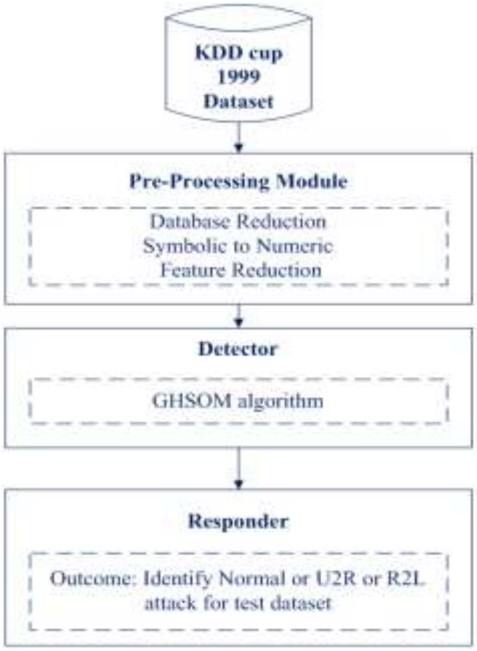

Figure 1: Proposed Model

In our pre-processing, there are three steps which can be followed in any order. First, as neural network do not accept symbolic data, they are converted into numeric data and it is called encoding. Secondly, our model is for only Normal, U2R and R2L attacks. So Database Reduction is done where the records of DoS and Probe are prune from KDD cup 1999 dataset. Lastly, for improvement of detection rate of U2R and R2L attacks our model uses 16 relevant features which are calculated on basis of information gain [22]. The 16 features of KDD Cup 1999 dataset which are used are following: 1, 3, 4, 5, 6, 9, 10, 11, 14, 16, 18, 23, 24, 26, 36 and 39.

Now, for classification of attacks our model uses clustering neural network model i.e. GHSOM algorithm. In GHSOM algorithm there are seven steps [7]. The following steps are explained in detail, how the GHSOM grid dynamically grows and is able to detect U2R and R2L attacks efficiently.

Step 1: Orientation of neuron grid and initialization of parameter α_1 and α_2.

Initially in GHSOM the neuron grid size is set to 2 x 2 and parameters values are initialized. The parameter α_1 is responsible for the horizontal growth of neuron grid while parameter α_2 is responsible for the vertical growth of neuron grid. The value of parameter $\alpha_1$and α_2 should be, between 0 and 1 where $1 > \alpha1 >> \alpha2 > 0$.

Step 2: Initialization of weight vector.

For parent layer, initialization of weight vector for 2 x 2 neuron grid is done randomly. For the growing grid, the new neurons which are introduced into the grid are initialized by averaging the weights of neighbour neurons. While for lower sub-layer, weight vectors of neurons are initialized by average of their parent unit's weight vector.

Step 3: Training of map

A random input is selected from dataset; then Euclidean distance using equation (1) is calculated between input vector and each weight vector. The weight vector neuron with minimum distances is considered as Best Matching Unit (BMU). Once the BMU is obtain the Neighbourhood function is calculated using equation (2) and weights of the neurons are update till tmax or convergence using equation(3).

$$\|x(t) - w_c(t)\| \leq \|x(t) - w_i(t)\|$$
$$\text{i.e.}\quad c = \min_i\{\,\|x(t) - w_i\,\|\} \qquad (1)$$

$$N_{c(x),i} = \exp(-\|w_c - w_i\| \,/\, 2\sigma(t)^2\ ) \qquad (2)$$

$$w_i(t+1) = w_i(t) + \delta(t)[\,x(t) - w_i\,(t)]N_{c(x),i} \qquad (3)$$

Step 4 and 5: Calculating quantization error

The mean quantization error (mqei) and Map's Quantization Error (MQEm) is calculated using equation (4) and equation (5) respectively.

$$mqe_i = (\,1 \,/\, \prod_c)\sum\|w_i - x_i\| \qquad (4)$$

$$MQE_m = 1 \,/\, U_m \sum mqe_i \qquad (5)$$

Step 6: Horizontal growth

If $MQE_m \geq \alpha_1 * mqe_u$ is satisfied then a new row or column of neurons is insert between the error neuron (e) and dissimilar neuron (d).

Step 7: Vertical growth

If $mqe_i \leq \alpha_2 * mqe_0$ is not satisfied then a new neuron sub-layer is introduced.

After training is completed using GHSOM algorithm, cluster labelling is required. In cluster labelling, if a neuron projects all input vector as normal then is it labelled as normal. If a neuron projects a particular attack then it is labelled accordingly. If a neurons projects to both normal and attacks then majority is taken and labelled accordingly. If no input vector is projected then neuron is labelled as empty cluster.

For testing, the Euclidean Distance is calculated for input vector and weight vector using equation (1), and BMU is selected. The BMU neuron has particular label. That particular label is then assigned to the input vector of test dataset.

In our model performance measure is done on the basis of True Positive, True Negative, False Positive, False Negative, Accuracy, Precision and Detection rate of individual attacks. Following are the equation used for performance analysis.

1) $\text{Positive (TP)} = \frac{\text{Correct Detected Attacks}}{\text{Total no.of Attacks}}$

2) $\text{False Positive (FP)} = \frac{\text{No.of Normal Detected as Attack}}{\text{Total no.of Normal}}$

3) $\text{True Negative (TN)} = \frac{\text{Correct Detected Normal}}{\text{Total no.of Normal}}$

4) $\text{False Negative (FN)} = \frac{\text{No.of Intrusion Detected as Normal}}{\text{Total no.of Attacks}}$

5) $\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$

6) $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$

7) $\text{Detection Rate} = \frac{\text{Correctly Detected Attacks}}{\text{Total number of Attacks}}$

EMPIRICAL RESULTS

In implementation we have carried out two experiments. The first experiment was performed to find the optimized value of parameter α_1 and α_2. And the second experiment was performed using various features and dataset for performance evaluation.

In pre-processing of KDD cup 1999 dataset we have reduce the training dataset from 4,94,021 records to 98,456 records and for testing dataset from 3,11,029 records to 77010 records.

Table 1: Accuracy and Precision for different value of α_1 ***and*** α_2

α_1	0.1		0.2		0.3		0.4	
α_2	Accuracy (%)	Precision (%)	Accuracy (%)	Precision (%)	Accuracy (%)	Precision (%)	Accuracy (%)	Precision (%)
0.01	95.43	95.37	98.25	97.42	98.04	96.14	97.40	96.23
0.02	95.72	96.03	96.83	97.55	98.54	98.67	98.15	98.43
0.03	97.28	95.07	96.53	96.16	97.24	97.92	98.83	99.14
0.04	96.25	96.92	95.87	95.53	97.30	98.20	98.10	96.56
0.05	98.09	97.20	97.17	95.70	96.54	97.71	95.42	95.81
0.06	98.74	97.88	97.25	99.05	97.80	96.15	96.33	96.46
0.07	95.61	96.16	97.68	97.77	96.70	96.43	97.81	97.77
0.08	96.33	96.76	96.49	96.86	98.72	99.35	97.68	96.38
0.09	97.71	98.25	98.02	97.74	97.30	96.99	98.18	98.38

Experiment-1: Optimization of Parameter

This experiment it carried out to find the optimized value for α_1 and α_2 . So that, further experiments where done using the optimized value of α_1 and α_2.

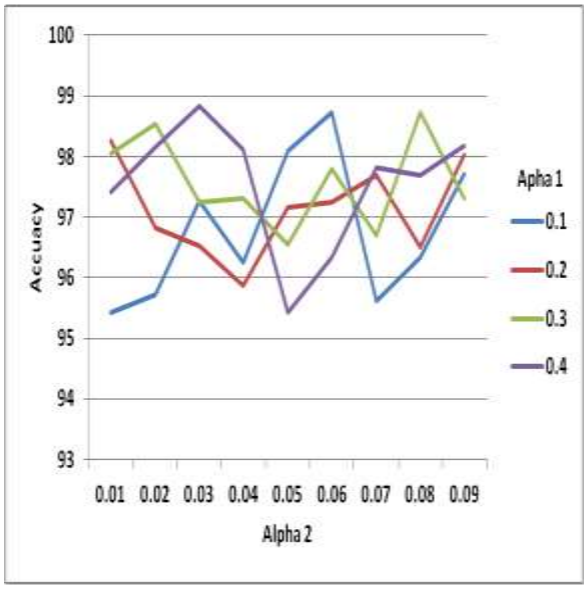

Figure 2: Accuracy for different values of α_1 and α_2

The table 1 show the exact values of accuracy and precision for different values of α_1 and α_2 .The accuracy is highest at $\alpha_1 = 0.4$ and $\alpha_2 = 0.03$ as shown in figure 2 and table 3. Hence, further experimentation is done with optimized value of α_1 and α_2 i.e. $\alpha_1 = 0.4$ and $\alpha_2 = 0.03$.

Experiment 2: Performances Evaluation with different feature and different dataset.

Experiment 2 is conducted under four different scenarios:

1. With 16 features and prune dataset having records of Normal, U2R and R2L attacks.
2. With 22 features and prune dataset having records of Normal, U2R and R2L attacks.
3. With 16 features and whole KDD cup dataset.
4. With 22 features and whole KDD cup dataset.

In scenario 1 and 3, the 16 features selected are relevant to U2R and R2L attacks while in scenario 2 and 4, the 22 features selected are relevant to all attacks. In scenario 1 and 2, the experimentation is done with prune having records of Normal, U2R and R2L attacks. From prune train dataset randomly 50000 records are used for training and all records of prune test dataset are used for testing. While in scenario 3 and 4, the experimentation is done with KDD cup 1999 train dataset random selecting 250000 records for training network and whole test dataset for testing.

Table 2: Experiment with different scenarios and comparison for it with detection rate of various attacks

Scenario	Normal (%)	U2R (%)	R2L (%)	DoS (%)	Probe (%)
1	90.12	70.77	74.55	NA	NA
2	95.58	60.12	64.10	NA	NA
3	95.64	66.20	65.70	73.49	71.82
4	91.48	64.48	60.76	80.28	82.10

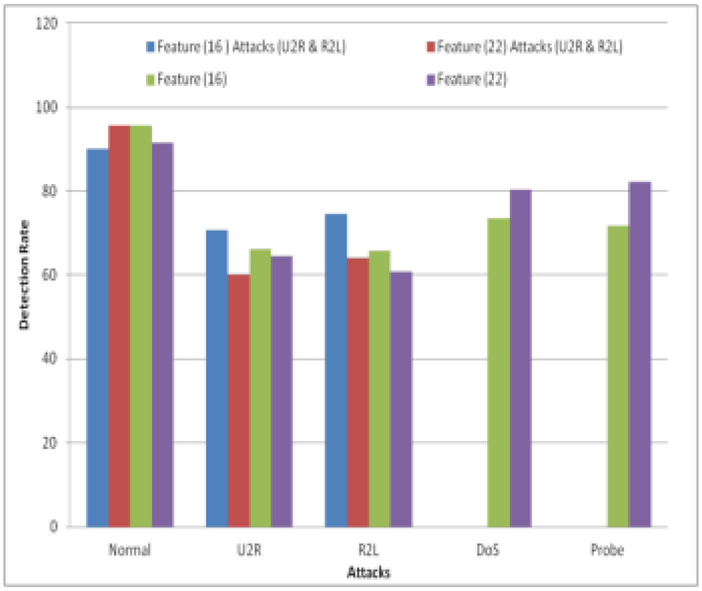

Figure 3: Detection rate of various attacks in various scenarios

From the table 2 and figure 3, it can concluded that, in the scenario 1 the detection rate of U2R and R2L attacks is highest which our proposed work. And the detection rate obtained for U2R and R2L attacks is 70.77% and 74.55% respectively.

COMPARISION OF GHSOM AND SOM

Now, comparison of the empirical result of GHSOM approach is done with SOM [13], as the objective of both works is same, to improve the detection rate of U2R and R2L attacks.

Table 3: Comparison of detection rate in GHSOM approach with SOM approach

	Normal (%)	U2R (%)	R2L (%)
GHSOM (Proposed Work)	90.12	70.77	74.55
SOM[13]	82.36	13.80	0

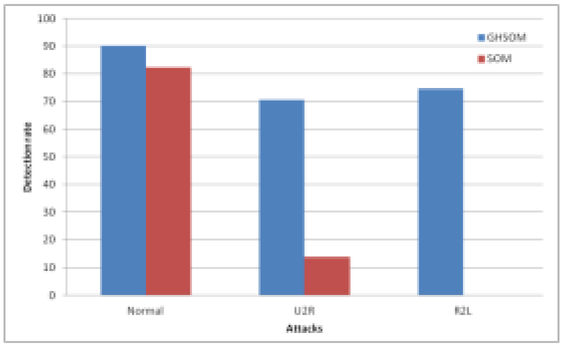

Figure 4: Comparison of GHSOM with SOM approach

The table 3 and figure 4, shows the comparison between results of the detection rate of Normal, U2R and R2L attacks in SOM and GHSOM algorithm. There is a significant increase in the detection rate of U2R and R2L attacks which can be seen in figure 4. Approximate 75% increase is there, in the detection rate of U2R and R2L attacks using GHSOM compared to SOM.

CONCLUSION

To improve the detection rate of U2R and R2L attacks our model used GHSOM algorithm and 16 relevant features to U2R and R2L attacks. From our survey, the existing IDS based on SOM face difficulties of more computation time and poor detection rate of U2R and R2L attacks. To overcome these difficulties, our model used GHSOM approach to deal with the more computation time and relevant feature to U2R and R2L attacks to enhance the performance. Our empirical result shows that, there is significant improvement in results. As approximately there is 75% increase in the detection rate of U2R and R2L attacks using our model compared to SOM approach.

DECLARATION

The content of this paper is written by Author 1(Kruti Choksi) while Author 2(Prof. Bhavin Shah) had guided the work and Author 3(Asst. Prof Ompriya Kale) has reviewed this paper. Hence Author 1 is responsible for the content and issues related with plagiarism.

REFERENCES

[1] Kumar, Gulshan, Krishan Kumar, and Monika Sachdeva. "The use of artificial intelligence based techniques for intrusion detection: a review." *Artificial Intelligence Review 34.4 (2010): 369-387.*

[2] Bashir, Uzair, and Manzoor Chachoo. "Intrusion detection and prevention system: Challenges & opportunities." *Computing for Sustainable Global Development (INDIACom), 2014 International Conference on. IEEE, 2014.*

[3] Alsulaiman, Mansour M., et al. "Intrusion Detection System using Self-Organizing Maps." *Network and System Security, 2009. NSS'09. Third International Conference on. IEEE, 2009.*

[4] Pachghare, V. K., Parag Kulkarni, and Deven M. Nikam. "Intrusion detection system using self organizing maps." *Intelligent Agent & Multi-Agent Systems, 2009. IAMA 2009. International Conference on. IEEE, 2009.*

[5] Wang, Chun-dong, He-feng Yu, and Huai-bin Wang. "*Grey self-organizing map based intrusion detection." Optoelectronics Letters 5 (2009): 64-68.*

[6] Ippoliti, Dennis, and Xiaobo Zhou. "An adaptive growing hierarchical self organizing map for network intrusion detection." *Computer Communications and Networks (ICCCN), 2010 Proceedings of 19th International Conference on. IEEE, 2010*.

[7] Mansour, Nashat, Maya I. Chehab, and Ahmad Faour. "*Filtering intrusion detection alarms." Cluster Computing 13.1 (2010): 19-29.*

[8] Salem, Maher, and Ulrich Buehler. "An Enhanced GHSOM for IDS." *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on. IEEE, 2013*.

[9] Ippoliti, Dennis, and Xiaobo Zhou. "A-GHSOM: An adaptive growing hierarchical self organizing map for network anomaly detection." *Journal of Parallel and Distributed Computing 72.12 (2012): 1576-1590*.

[10] Palomo, Esteban J., et al. "A new GHSOM Model applied to network security. *".Artificial Neural Networks-ICANN 2008. Springer Berlin Heidelberg, 2008. 680-689.*

[11] Ortiz, Andres, et al. "Improving Network Intrusion Detection with Growing Hierarchical Self-Organizing Maps." University of De La Plata, Argentina (2011).

[12] Gunes Kayacik, H., A. Nur Zincir-Heywood, and Malcolm I. Heywood. "A hierarchical SOM-based intrusion detection system." Engineering Applications of Artificial Intelligence 20.4 (2007): 439-451.

[13] Wilson, Ryan, and Charlie Obimbo. "Self-organizing feature maps for user-to-root and remote-to-local network intrusion detection on the KDD cup 1999 dataset." Internet Security (WorldCIS), 2011 World Congress on. IEEE, 2011.

[14] Bahrololum, M., E. Salahi, and M. Khaleghi. "Anomaly intrusion detection design using hybrid of unsupervised and supervised neural network."International Journal of Computer Networks & Communications (IJCNC) 1.2 (2009): 26-33.

[15] Kohonen, Teuvo. "The self-organizing map." Proceedings of the IEEE 78.9 (1990): 1464-1480.

[16] Ibrahim, Laheeb M., Dujan T. Basheer, and Mahmod S. Mahmod. "A Comparison Study For Intrusion Database (Kdd99, Nsl-Kdd) Based On Self Organization Map (SOM) Artificial Neural Network." Journal of Engineering Science and Technology 8.1 (2013): 107-119.

[17] Huang, Shin-Ying, and Yennun Huang. "Network forensic analysis using growing hierarchical SOM." Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on. IEEE, 2013.

[18] Huang, Shin-Ying, and Yen-Nun Huang. "Network traffic anomaly detection based on growing hierarchical SOM." Dependable Systems and Networks (DSN), 2013 43rd Annual IEEE/IFIP International Conference on. IEEE, 2013.

[19] Zolotukhin, Mikhail, T. Hamalainen, and Antti Juvonen. "Online anomaly detection by using N-gram model and growing hierarchical self-organizing maps." Wireless Communications and Mobile Computing Conference (IWCMC), 2012 8th International. IEEE, 2012.

[20] Tesfahun, Abebe, and D. Lalitha Bhaskari. "Intrusion Detection Using Random Forest Classifier with SMOTE and Feature Reduction." Cloud & Ubiquitous Computing & Emerging Technologies (CUBE), 2013 International Conference on. IEEE,2013.

[21] Kayacik, H. Günes, A. Nur Zincir-Heywood, and Malcolm I. Heywood. "Selecting features for intrusion detection: A feature relevance analysis on KDD 99 intrusion detection datasets." Proceedings of the third annual conference on privacy, security and trust. 2005.

[22] Olusola, Adetunmbi A., Adeola S. Oladele, and Daramola O. Abosede. "Analysis of KDD'99 Intrusion detection dataset for selection of relevance features." Proceedings of the World Congress on Engineering and Computer Science. Vol. 1. 2010.

[23] KDD Cup 1999 Data.
http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[24] Kruti Choksi, Prof. Bhavin Shah, Asst. Prof. Ompriya Kale, "Intrusion Detection System using Self Organizing Map: A Survey"Vol. 4 - Issue 12 (December - 2014), International Journal of Engineering Research and Applications (IJERA).

[25] Shah, Bhavin, and Bhushan H. Trivedi. "Artificial neural network based intrusion detection system: A survey." International Journal of Computer Applications 39.6 (2012).