SUPERVISED LEARNING VECTOR QUANTIZATION AND ENHANCED BACKPROPAGATION NEURAL NETWORK FOR INTRUSION DETECTION

K.Juliana Gnanaselvi [1], Dr.V.Shyamala Susan[2]

1.Asst Professor, Dept of Computer Science, Rathinam college of Arts and Science, Coimbatore

2. Asst Professor, Dept of Computer Science, A.P.C.Mahalaxmi College for Women,Thoothukudi

ABSTRACT

Network-based computer systems have become the target of intrusions by adversaries. Intrusion Detection System (IDS) tries to notice computer attacks by inspecting various data records observed in processes on the network. This paper presents an intrusion detection system models, using Supervised Learning Vector Quantization (SLVQ) and an enhanced Neural Network algorithm with Backpropagation (NNBP). A Supervised Learning Vector Quantization as the first stage of classification was trained to detect intrusions. In the next stage the algorithm continues with identifying minimal number of hidden units in the single hidden layer; then additional units are added to the hidden layer one at a time to enhance the accuracy of the network and to get an optimal size of a neural network. An optimal learning factor was derived to speed up the convergence of the Neural Network algorithm with Backpropagation performance. The evaluations were performed using the NSL-KDD99 network anomaly intrusion detection dataset. The experiments results demonstrate that the proposed system (SLVQ_NNBP) has a detection rate about 98.06% with a false negative rate of 3%.

Keywords: *Intrusion Detection System, Supervised Learning Vector Quantization, enhanced neural network, backpropagation*

1. INTRODUCTION

Protecting the systems against attack and intrusion is a critical task, as many companies and government agencies rely on computer network [1]. IDS have become the major issue of network security. The two intrusion detection techniques are misuse detection and anomaly detection. Misuse detection systems, use patterns of recognized attacks or weak spots of the system to match and identify well-known intrusions. In anomaly detection systems, flag observed events that deviate significantly from the recognized normal usage profiles as anomalies, that is, possible intrusions. An anomaly detection technique is an effective technique because priori knowledge about specific intrusions is not required. However, anomaly detection systems mostly generate more false alarms than misuse detection systems as an anomaly can just be a new normal behavior [2].

Neural networks on the other hand, are powerful tool in multiple classifications, especially when used in applications when formal analysis are difficult to identify, such as pattern recognition, nonlinear system identification, and control [3]. Neural networks are able to work with indefinite and incomplete data because of their generalization feature. They can also recognize patterns not presented during a learning phase. Thus the neural networks are a good solution for detecting a well- known attack, which has been modified by an intruder. In such case, traditional IDS, based on the signatures of attacks or expert rules, may not be able to identify the new version of this attack [4].

Backpropagation algorithm is the widely used learning algorithm to train multiplayer feedforward network and applied for applications such as character recognition, image processing, pattern classification etc. The network must be built before we train an artificial neural network (ANN). All the nodes in the input layer, output layer and the hidden layer must be defined. Nowadays, many researches have been done on algorithms that dynamically build neural networks for solving pattern classification problems. In this proposed research we proposed an algorithm, which can add nodes in the single hidden layer during the training period and can build an ANN with its minimal size, which can classify intruder with acceptable efficiency.

2. RELATED WORKS

Depren et al. (2005) [5] proposed an intelligent IDS for anomaly detection system with the aid of Self Organizing Map (SOM) to model the normal behavior. This model used powerful unsupervised SOM which results a low false positive rate, but on the other hand the system didn't classify the records into 5 classes. Ahmad, Swati & Mohsin (2007) [6] used resilient backpropagation for intrusion detection. The ANN architecture has an input and output layers and two hidden layer, with 41, 14, 9, and 2 neurons respectively. The proposed system had a very good accuracy rate but on the other hand they have used 2 hidden layers which are not necessary particularly if the neural network parameters were selected optimally. Naoum, Abid and Al-Sultani (2012) [7] proposed a hybrid intrusion detection system based on k-Nearest Neighbor and an enhanced resilient backpropagation artificial neural network. An optimal learning factor was derived to speed up the convergence of the enhanced resilient backpropagation. k-Nearest Neighbor implementation used first normal form instead of Euclidean distance and they have used the first nearest neighbor where k equals 1. The enhanced resilient backpropagation neural network trained by means of an optimal number of hidden layers and neurons; thus it was trained with only one hidden layer and 34 hidden neurons. The evaluation was performed on the NSL-KDD99 anomaly intrusion detection dataset. The proposed system has a classification rate (5 classes) of 97.2% with false negative rate of about 1%.

3. PROPOSED SYSTEM

This research work tries to classify intrusions, using SLVQ and NNBP. The system is tested and evaluated using the NSL-KDD dataset. The proposed system is divided into five phases: environment phase, dataset features and pre-processing phase, SLVQ phase, enhanced NNBP phase and testing the system phase.

3.1 The Environment Phase

This unit presents records from NSL KDD99 dataset [8]. This data set is divided into two subsets namely training subset and testing subset. The NSL KDD dataset includes a wide variety of intrusions together with normal activities simulated in a military network environment. NSL KDD records belong to one of the following five categories: Normal, DoS (denial of service), R2L (root to local), U2R (user to root) and Probing (surveillance). There are 41 features columns and they are either symbolic or continuous.

3.2 Data Pre-processing Phase

The data from the environment phase will be processed before entering the classification unit. Feature columns are processed at 2 steps as transformation and standardization.

1. Transformation: Symbolic columns are distorted to numeric values using transformation table for each column. Table 1 demonstrate the transformation table for flag feature column.

Table 1 Flag Column Feature Transformation Table

Flag-4	No
OTH	1
REJ	2
RSTO	3
RSTO 0	4
RSTR	5
S0	6
S1	7

S2	8
S3	9
SF	10
SH	11

Label column (column 42) contains either normal or the sub-type attack label. Transforming this column was done in two steps. First the sub-attack type was represented with the main attack type, and then the main attack type was transformed to numeric using 5 columns, each class is represented with value one using one column. Table 2 represents the customization transformation for the main classes.

Table 2 Label Transformation Table

Label -42	Column1	Column2	Coulmn3	Column4	Column5
Normal	**1**	**0**	**0**	**0**	**0**
DoS	**0**	**1**	**0**	**0**	**0**
U2R	**0**	**0**	**1**	**0**	**0**
R2L	**0**	**0**	**0**	**1**	**0**
Prob.	**0**	**0**	**0**	**0**	**1**

2. Standardization: Training subset matrix is processed by mapping each row's means to 0 and standard deviations to 1. It's important to indicate that the main testing dataset also should be standardized using the mean and the variance of the training dataset before performing the simulation.

3.3 Supervised Learning Vector Quantization phase

The SLVQ is a nearest neighbour pattern classifier based on competitive learning. Here this SLVQ is trained to identify intrusions in the first stage. Kohonen originally suggested it. The basic architecture of SLVQ neural network is shown in Fig.1.

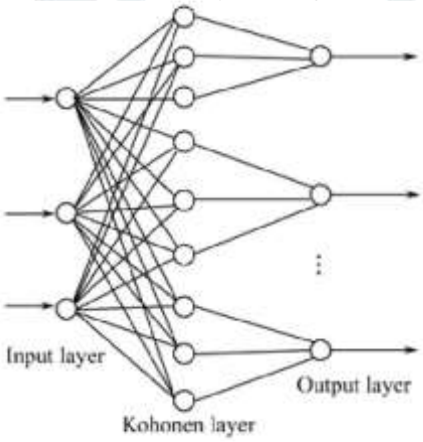

Fig.1 The architecture of LVQ neural network

In Fig.1, LVQ network contains an input layer, a Kohonen layer which learns and performs the classification, and an output layer. The input layer contains one node for each input feature; the Kohonen layer comprises of equal numbers of nodes for each class; in the output layer, each output node signifies a particular class. Its main indication is to divide the input space Rn into a number of distinct regions, called decision regions, and for each region one reference vector is assigned. Classification is performed based on the vicinity of the input vector X to the reference vectors; X will be classified as the label of its nearest neighbour among reference vectors. During the training, the reference vectors and thus the borders of decision regions are adjusted through an iterative process. Subsequently, the LVQ is ready to classify the testing dataset. LVQ will classify the dataset into five classes (Normal, DoS, U2R, R2L and Prob). Then the results of LVQ will be combined later with the results of the neural network trained using the NNBP classifier to provide maximum classification rates.

3.3 Enhanced neural network with backpropagation phase

One of the difficulties with the traditional backpropagation algorithm is the decision of the number of neurons in the hidden layer within a network. To overcome this problem the enhanced NNBP for feedforward networks may be used, which constructs the network during training. Hence an optimal number of neurons can be generated in the hidden layer to attain a satisfactory level of efficiency for a particular problem. In addition to applying the early stopping method of training using cross-validation we can also train the network in a relatively short estimation period (training period). In the construction algorithm proposed by Rudy Setiono and Huan Liu they have defined the stopping condition of the training by classifying all the input patterns. It means that while the efficiency is 100%, the training will stop. But in most cases with the benchmarking classification problems 100% efficiency may not be attained. In such case we adopt a new algorithm for pattern classification that defines the stopping condition by the acceptance of efficiency level. Also we made that the desired efficiency on the test sets may not be achieved even though the mean square error on training set is minimum. These concerns motivated to propose an algorithm that will combine the learning rule of backpropagation algorithm to update weights of the network and the construction algorithm to construct the network dynamically and also consider the efficiency factor as a determinant of the training process.

The following steps are followed to build and train a network [9];

1. Create an initial neural network with number of hidden unit *h* = 1. Set all the initial weights of the network randomly within a certain range.
2. Train the network on training set by using a training algorithm for a certain number of epochs that minimizes the error function.
3. If the error function ξ_{av} on validation set is acceptable and, at this position, the network classifies desired number of patterns on test set that leads the efficiency *E* to be acceptable then *stop.*
4. Add one hidden unit to hidden layer. Randomly initialize the weights of the arcs connecting this new hidden unit with input nodes and output unit(s).
   Set *h* = *h* + 1 and go to *step 2.*

To provide maximum generalization, we started with only one hidden layer using different number of hidden neurons iteratively. Here iterative process is used because high number of hidden neurons will lead to over- fitting problem, where the neural network will not be able to classify new records. Generally if there are no good results then a second layer can be added to improve the neural performance. Experiments have shown that when using only one hidden layer with 32 hidden neurons, the enhanced NNBP performance gave the best classification rate.

3.4 Testing the hybrid system (SLVQ_NNBP) Phase

In this phase, testing dataset will be classified by both SLVQ and the NNBP which was trained during the training phase using the best number of hidden neurons and layers. The propsed system is evaluated by calculating the Detection Rate (DR), False Positive Rate (FPR) etc.

4. EXPERIMENT RESULTS

In this paper a hybrid system of SLVQ and the NNBP was trained to detect intrusions using NSL-KDD99 dataset. Testing set contains some attacks that it is not represented in the training set. In short, intrusions are generally classified into several categories Attack types that are classified as:

- Denial of service (DoS)
- Probe (PRB)

- Remote to login (R2L)
- User to root (U2R)

The NNBP as the second classifier will be used also to classify the testing dataset into 5 classes. After combining the results of both classifiers the class detection rate of the hybrid (SLVQ_NNBP) is shown in table 3:

Table 3 Hybrid (SLVQ_NNBP) Detection Rate

Testing(Labeled) Datasets	*Class Size*	*Detected Size*	*Detection Rate*
Normal	1000	923	92.3%
DoS	2200	2155	97.9%
U2R	37	30	81.0%
R2L	2200	2127	96.6%
Prob.	2200	2199	99.9%
Total	7637	7531	98.6%

False Positive Rate, False Negative Rate, Recall, and Precision metrics are used to estimate the performance of learning algorithms. Table 4 shows the values of these metrics for the hybrid system (SLVQ_NNBP):

Table 4 Hybrid System (SLVQ_NNBP) Evaluation Metrics

Testing(Labeled) Datasets	*Percentage*
Recall	97%
Precision	99%
False Negative Rate	3%
False Positive Rate	9%

Table 5 shows the comparison between the proposed SLVQ_ NNBP and the SLVQ_kNN

Table 5 Hybrid (SLVQ_* NNBP) vs. Hybrid (SLVQ_kNN)*

Dataset	**Hybrid (SLVQ_NNBP)**	**%**	**Hybrid (SLVQ_kNN)**	**%**
Normal	923 – 1000	92.3%	938 – 1000	93%
DoS	2155 – 2200	97.9%	1187 -1200	98%
U2R	37-30	81.0%	30 – 37	81%
R2L	2127 – 2200	96.6%	194 – 500	39%
Prob.	2199 – 2200	99.9%	1157 – 1200	96%
All	7531 – 7637	98.6%	3506 – 3937	89%

5. CONCLUSION

In this paper SLVQ and an enhanced NNBP were trained to detect intrusion. The main issue in SLVQ training is, it needs a long time to be trained especially when associating to other networks such as Multilayer Perceptron or Self Organizing Maps. The experimental results shows that the SLVQ_ NNBP had better results than SLVQ_kNN. User to Root as a low-frequent has the lowest detection rate among other classes. This is because the leaning sample size is too small compared to high-frequent attacks, hence it makes SLVQ_ NNBP not easy to learn the characters of these attacks and therefore detection precision is much lower.

6. REFERENCES

[1] Cannady, J. (1998). Artificial Neural Networks for Misuse Detection. *National Information Systems Security Conference*. Retrieved October 18, 2011, from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.39.5179

[2] LEE, W. & STOLFO, S (2000). A Framework for Constructing Features and Models for Intrusion Detection Systems. ACM Transactions on Information and System Security, 3 (4).

[3] Sammany, M., Sharawi, M., El-Beltagy, M. & Saroit, I. (2007). Artificial Neural Networks Architecture For Intrusion Detection Systems and Classification of Attacks. Faculty of Computers and Information Cairo University. Retrieved October 18, 2011, from http://infos2007.fci.cu.edu.eg/Computational%20Intelligence/07177.pdf

[4] Kukiełka, P. & Kotulski, Z. (2010). Adaptation of the neural network-based IDS to new attacks detection. arXiv - Cornell University. Retrieved October 26, 2011, from http://arxiv.org/ftp/arxiv/papers/1009/1009.2406.pdf

[5] Depren, O., Murat, T., Anarim, E. & Ciliz, M. (2005). An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks. *Expert Systems with Applications, Elsevier*,713-722. Retrieved October 20, 2011, from http://www.ft.unicamp.br/RedesComplexas/downloads/An_intelligent_intrusion_detection_system_for_anomal y_and_misuse_detection_in_computer_networks.pdf

[6] Ahmad, I., Swati, S. & Mohsin, S. (2007). Intrusions Detection Mechanism by Resilient Back Propagation (RPROP). *European Journal of Scientific Research, EuroJournals Publishing, Inc*, *17*, 523-531. Retrieved November 2, 2011, from http://www.eurojournals.com/ejsr%2017%204.pdf

[7] Naoum,R. Abid,N. & Al-Sultani,Z. (2012). A Hybrid Intrusion Detection System Based on Enhanced Resilient Backpropagation Artificial Neural Network and K-Nearest Neighbor Classifier. International Journal of Academic Research IJAR, 4 (2). Retrieved from http://www.ijar.lit.az/en.php?go=march2012

[8] Information Security Center of eXcellence (ISCX), The NSL-KDD Data Set, 2009. Retrieved October 26, 2011, from http://www.iscx.ca/NSL-KDD/

[9] Rudy Setiono and Huan Liu, “Improving Backpropagation Learning with Feature selection”, Appears in Applied Intelligence, Vol. 6, No. 2, 1996, pp. 129-140.