2011 International Conference on Telecommunication Technology and Applications
Proc .of CSIT vol.5 (2011) © (2011) IACSIT Press, Singapore

Feature Subset Selection for Network Intrusion Detection Mechanism Using Genetic Eigen Vectors

Iftikhar Ahmad[1], Azween B Abdulah[2], Abdullah S Alghamdi[1], Khaled Alnfajan[1]
Muhammad Hussain[3]

[1] Department of Software Engineering, College of Computer and Information Sciences, P.O. Box 51178, Riyadh 11543, King Saud University, Saudi Arabia.

[2] Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia.

[3] Department of Computer Science, College of Computer and Information Sciences, P.O. Box 51178, Riyadh 11543, King Saud University, Saudi Arabia.

{wattoohu@gmail.com}

Abstract. Network Intrusions are critical issues in computer and network systems. Several intrusion detection approaches be present to resolve these severe problems but the major problem is performance. To increase performance, it is significant to increase the detection rates and reduce false alarm rates in the area of intrusion detection. The recent approaches use Principal Component Analysis (PCA) to project features space to principal feature space and select features corresponding to the highest eigenvalues, but the features corresponding to the highest eigenvalues may not have the optimal sensitivity for the classifier due to ignoring many sensitive features. Instead of using traditional approach of selecting features with the highest eigenvalues such as PCA, we applied a Genetic Algorithm (GA) to search the principal feature space for genetic eigenvectors that offers a subset of features with optimal sensitivity and the highest discriminatory power. Therefore, in this research, a mechanism for optimal features subset selection is proposed to overcome performance issues using PCA, GA and Multilayer Perceptron (MLP). The KDD-cup dataset is used that is a benchmark for evaluating the security detection mechanisms. The MLP is used for classification purpose. The performance of this approach is addresses. Consequently, this method provides optimal intrusion detection mechanism which is capable to minimize amount of features and maximize the detection rates.

Keywords: KDD-cup dataset, Principal Component Analysis (PCA), Multilayer Perceptron (MLP), Genetic Algorithm (GA) , Detection Rates (DRs), False Alarms (FAs),False Positives (FPs), False Negatives (FNs), True Positives (TPs), and True Negatives (TNs)

1. Introduction

The existing approaches of intrusion detection have focused on the issues of feature extraction and classification. However, comparatively less concentration has been given to the critical matter of feature selection. The foremost trend in feature extraction has been representing the data in to another feature space (the PCA space) using principal component analysis (PCA). In this method of selecting features on the basis of highest eigenvectors is not appropriate because the features corresponding to the highest eigenvalues may not have the optimal sensitivity for the classifier due to ignoring many sensitive features. So, there must be an effective scheme to select an appropriate set of features in the PCA space. This will leads the classifier to work in an efficient way and increases the overall performance of the intrusion analysis engine. Because, the redundant and irrelevant features increases overheads as well as confuses the classifier. Therefore, in this paper, we argue that feature selection is an important problem in intrusion detection and demonstrate that

genetic algorithms (GAs) provide a simple, general, and powerful framework for selecting good subsets of features that improve detection rates [1]. Further, we considered PCA for features transformation and MLP for classification. The goal is searching the PCA space using GA to select a subset of principal components. This is in contrast to traditional methods selecting some percentage of the top principal components to represent the target concept, independently of the classification task. We have tested the proposed framework on intrusion detection. Our experimental results illustrate significant performance improvements. Several approaches have been reported in the area of intrusion detection but the main focus is on classification. In [2], PCA is applied for classification and neural networks are used for online computing. They selected 22 principal components as features subset selection to obtain the best performance. But there is a possibility to miss many important principal components having sensitive information for intrusion detection during selection phase.

In [3], the importance feature is determined based on the accuracy and the number of false positives of the system, with and without the feature. In other words, the feature selection of is "leave-one-out"; remove one feature from the original dataset, redo the experiment, then compare the new results with the original result, if any case of the described cases occurs. The feature is regarded as important; otherwise it is regarded as unimportant. Since there are 41 features suggested in the KDD-cup99, the experiment is repeated 41 times to ensure that each feature is either important or unimportant. This method involved complexity as well as overheads on huge dataset. In [4], the radial basis function (RBF) network is employed as a real-time pattern classification and the Elman network is employed to restore the memory of past events. They used full featured KDD cup dataset. This increases training and testing overheads on the system. In [5], PCA method is used to determine an optimal feature set. An appropriate feature set helps to build efficient decision model as well as to reduce the population of the feature set. Feature reduction will speed up the training and the testing process for the attack identification system considerably but this will be a compromise between training efficiency (few PCA components) and the accurate results (a large number of PCA components).

In [6], the fusions of Genetic Algorithm (GA) and Support Vector Machines (SVM) are described for optimization of both features and parameters for detection models. This method was capable to minimize amounts of features and maximize the detection rates but the problem is features uniformity. The features in original forms are not consistent so these must be transformed in new feature space in order to well organized form. Let us describe proposed model and its methodology for features subset selection.

2. Proposed Model

The proposed model consist of different parts; dataset used for experiments, feature transformation and organization, optimal feature subset selection, classification architectures, implementation, training and testing, and results comparison. The block diagram of proposed model is shown in the Figure 1.

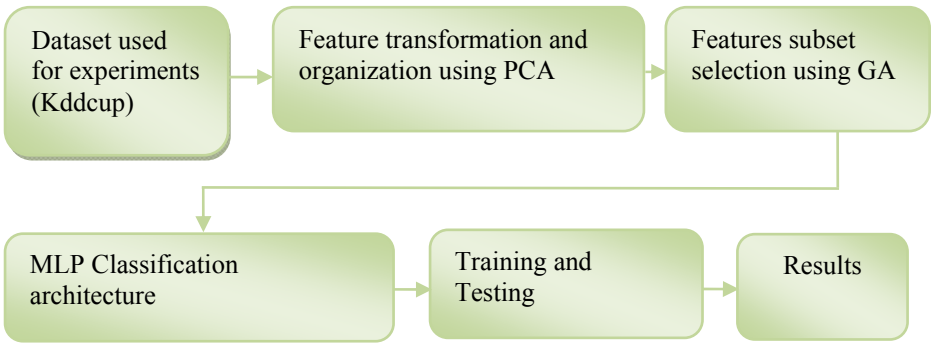

Fig 1: Proposed Model

2.1. Dataset used for Experiments

We used kddcup99 dataset for our experiments. The selection of this dataset is due to its standardization, content richness and it helps to evaluate our results with existing researches in the area of intrusion detection. We have already described this selected dataset in chapter 5. The raw dataset consists of 41 features.

$$x_1, x_2, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots x_n \qquad \textit{Where } n=41 \qquad (1)$$

2.2 Dataset pre-processing for Experiments

After selection of the dataset, first, we pre-processed on the raw dataset so that it can be given to the selected classifiers; MLP. The raw dataset is pre-processed in three ways; (i) discarding symbolic values, (ii) feature transformation and organization using PCA, and (iii) optimal features subset selection using GA.

2.2.1 *Discarding symbolic values*

In first step of pre-processing, we discarded three symbolic values (e.g. udp, private & SF) out of 41 features of the dataset. The resultant features are;

$$x_1, x_2, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots x_m \qquad \textit{Where } m=38 \qquad (2)$$

2.2.2 Feature transformation and organization

In second step of pre-processing, we applied PCA on 38 features of the dataset. Mostly, PCA is used for data reduction, but here, we used it for feature transformation into principal components feature space and then organized principal components in descending order.

$$pc_1 > pc_2 > pc_3 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots > pc_l \qquad \textit{Where } l=38 \qquad (3)$$

2.2.3 Feature Subset Selection

In third step of pre-processing, we applied genetic algorithm (GA) for optimal features subset selection from principal components search space. We used the fitness function shown below to combine the two terms:

$$fitness = 10^4 Accuracy + 0.5Zeros \qquad (4)$$

Where Accuracy corresponds to the classification accuracy on a validation set for a particular subset of principal components and Zeros corresponds to the number principal components not selected. The Accuracy term ranges roughly from 0.50 to 0.99, thus, the first term assumes values from 5000 to 9900. The Zeros term ranges from 0 to L − 1 where L is the length of the chromosome, thus, the second term assumes values from 0 to 37 (L=38).

2.3 Classification Architectures

A multilayer perceptron (MLP) is a feedforward neural network that maps sets of input data onto a set of appropriate output. Here, we used a MLP architecture consists of three layers; input, hidden and output. In this architecture, hidden layer and output layer consist of neurons (processing elements) and each neuron has a nonlinear activation function. The layers are fully connected from one layer to the next. MLP is an amendment of the standard linear perceptron, which can discriminate data that is not linearly separable. The architecture, we used here is shown in Figure 2.

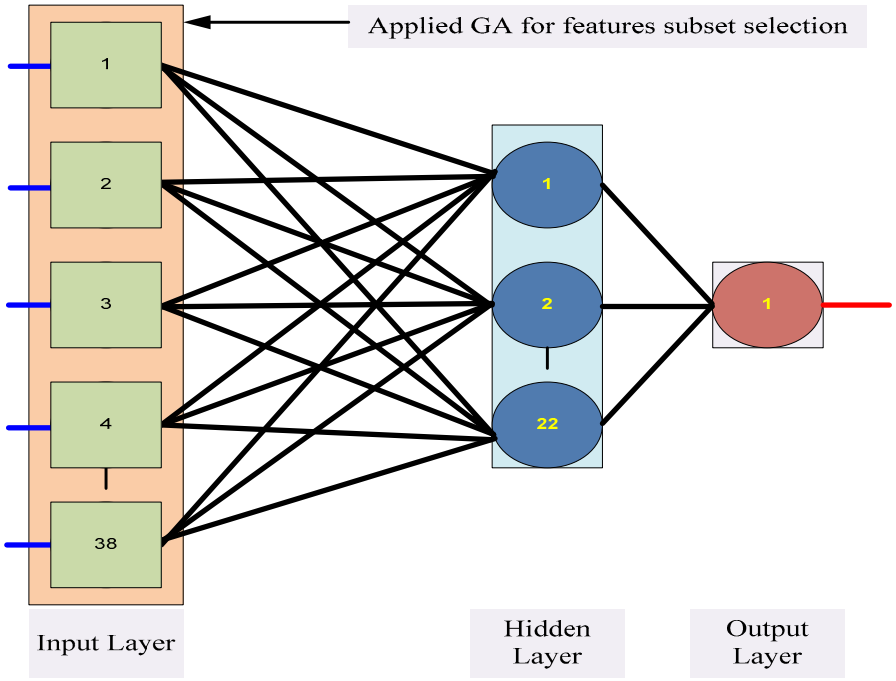

Fig 2: MLP Architecture for Network Intrusion Analysis

2.4 Training and Testing of the System

The aim of training is the adjustment of networks weights on base of the difference between the output produced by the system and the desired output. The training dataset consists of five thousand (5000) labeled connections (network packets with label as normal or intrusive) that are randomly selected from 20,000 connections. Further, we divide the training dataset (five thousand) into three parts; (i) cross validation dataset (1000), (ii) test dataset (1500) and (iii) training dataset (2500).

We used confusion matrix to verify the training. When the training is completed then weights of the system are frozen and performance of the system is evaluated. Testing the system involves two steps; (i) verification step, and (ii) generalization step. In verification step, system is tested against the data which are used in training. Aim of the verification step is to test how well trained system learned the training patterns in the training dataset. In generalization step, testing is conducted with data which is not used in training. Aim of the generalization step is to measure generalization ability of the trained network. We used a dataset of fifteen thousand (15,000) as a production dataset. We also tested our system performance on total dataset (20,000) that consist of both training dataset and production dataset.

2.5 Results

We performed three different experiments as shown in Table 1 and selected a subset of twelve features that indicates better performance as compared to others. Our aim is to select minimum features that produce optimal results in accuracy. This definitely impact on overall performance of the system.

Table 1 : Experimental Results

Experiment No.	Time	No. of selected PCs	No of non selected PCs	Accuracy	Fitness
1-MLP	72 hrs	12	26	0.99	9913
2-MLP	78 hrs	20	18	0.98	9808
3-MLP	83 hrs	27	11	0.99	9911

$$pc_2, pc_3, pc_9, pc_{11}, pc_{12}, pc_{15}, pc_{17}, pc_{18}, pc_{24}, pc_{27}, pc_{34}, pc_{36} \quad (5)$$

Twelve different principal components are selected using GA process

The features are reduced to 12 from the 41 raw features set. The above experiments show that optimal features increased accuracy, reduced training and computational overheads and simplified the architecture of intrusion analysis engine.

3. References

[1] Zehang Sun, George Bebis, Ronald Miller, *Object detection using feature subset selection*, Pattern Recognition, Volume 37, Issue 11, Nov. 2004, pp 2165-2176.

[2] Guisong Liu, Zhang Yi, Shangming Yang, *A hierarchical intrusion detection model based on the PCA neural networks*, Neurocomputing, Volume 70, Issues 7-9, pp 1561-1568.

[3] Shi-Jinn Horng, Ming-Yang Su, Yuan-Hsin Chen, Tzong-Wann Kao, Rong-Jian Chen, Jui-Lin Lai, Citra Dwi Perkasa, *A novel intrusion detection system based on hierarchical clustering and support vector machines*, Expert Systems with Applications, Volume 38, Issue 1, January 2010, pp306-313.

[4] Xiaojun Tong, Zhu Wang, Haining Yu, *A research using hybrid RBF/Elman neural networks for intrusion detection system secure model, Computer Physics Communications*, Volume 180, Issue 10, Oct. 2009, pp1795-1801.

[5] Gholam Reza Zargar and Peyman Kabiri, *Selection of Effective Network Parameters in Attacks for Intrusion Detection, Advances in Data Mining*. Applications and Theoretical Aspects, Lecture Notes in Computer Science, 2010, Volume 6171/2010,pp 643-652.

[6] Dong Seong Kim, Ha-Nam Nguyen, Syng-Yup Ohn and Jong Sou Park, *Fusions of GA and SVM for Anomaly Detection in Intrusion Detection System,* Advances in Neural Networks, Lecture Notes in Computer Science, 2005, Volume 3498/2005, pp415-420.