International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 4 (2015) pp. 8887-8901

http://www.ripublication.com

Multi Class Classifier Using Clustering Guided By PCA For Anomaly Based Risk Reduction

C.Kavitha[1] and Dr.K.Iyakutti[2]
[1]*Department of Computer Science, Pasumpon Muthuramalinga Thevar College, Madurai, Tamilnadu,India.*
[2]*Department of Physics & Nanotechnology, SRM University,Kattankulathur-603203, Chennai, Tamilnadu, India .*
[1]*kkavitha009@gmail.com*
[2]*iyakutti@gmail.com*

Abstract

The growth in the data consumption and the transfer of information is exploding. This explosion leads to the growth of the enterprise at the same time it paves the way for many malicious attacks for the information trapping. In order to reduce the risk for the information security attacks it is necessary for every enterprise to device a mechanism to prevent the data loss. This paper suggests a way to reduce the risk by the proactive anomaly detection. This paper deploy a method, which uses the PCA for the dimensionality reduction and which is further clustered to reduce the overhead of the classification for the anomaly detection. The results are summarized and tabulated. The results of the proposed approach are outperforming the existing methods. This shows the promising feature of the proposed approach.

Keywords: Anomaly detection, PCA, GA, Clustering, Classification,..

Introduction

As the communication technology grows faster and larger the attacks on the system which deploy the technologies are highly prone to the malicious attacks. These attacks are decidedly intentional to suppress the system from functioning. It creates the situation in which the attacks are to be overcome with various techniques to reduce the effect of the attacks. These attacks have a heavy impact on the functionality as well as the reputation of the enterprise deploying. This situation creates a constraint to detect and nullify the intentional attacks. As information management systems become more and more powerful and distributed, the number of threats grows and diversifies and there are many different ways to attack computers and networks [1].

Intrusion detection systems (IDS) are the famous techniques used to detect the attacks both in external form and in the internal form. IDS are defined as a process or device that analyzes system and network activity for unauthorized access and/or malicious activity [2]. IDS could be classified as misused based systems and anomaly based detections. Misused based detection catches the intrusions in terms of the characteristics of known attacks or system vulnerabilities. Anomaly based detection detect any action that significantly deviates from the normal behavior. Since the number of attacks and vulnerabilities are rising, and because of the inability of misuse detection functions to detect novel attacks that have no signatures yet [3], researchers are encourage to promote the intrusion detection mechanism to be able to detect novel attacks using anomaly detection [4].

The risk for the attacks could be reduced using the detection mechanisms. In this paper we propose a novel approach where the problem is encountered using the machine learning techniques to enhance the results of the detection mechanism. High dimension is the scalability factor associated with the large dataset. Feature selection process has its own disadvantages. This paper uses the dimensionality reduction using Principal components analysis (PCA) surge with the clustering process to come across the feature autonomy problem. The multi class classifier for the anomaly detection is built which in turn deploys genetic algorithms to increase the classification accuracy and downsize the error rate of the classifier.

The paper is organized in the following sections as section 2 talks about the background study of the concepts deployed in the paper, section 3 discuss on the existing methodology of the core problem. Problem formulation is thrashed out in the section 4. Section 5 deals with the proposed approach of this paper. The experimental details and the results obtained are depicted in the section 6. Discussion on the results is carried out in the section 7. Section 8 concludes the paper.

Background Study

This section briefly discuss on the concepts used in this paper. Anomaly detection, Dimensionality reduction, data mining is conferred in this section.

Risk reduction through Anomaly detection

The key to the value and effectiveness of anomaly based NIDS is that they can automatically infer attacks which are yet unknown, and therefore undetectable by signature based Intrusion detection systems. An anomaly detection technique generally consists of two different steps: the first step is called training phase wherein a normal traffic profile is generated; the second phase is called anomaly detection, wherein the learned profile is applied to the current traffic to look for any deviations. The anomaly detection problem can be considered as a two-class classification problem (normal versus abnormal) where samples of only one class (normal class) are used for training. A number of anomaly detection mechanisms have been proposed recently to detect such deviations, which can be categorized into [5]

1. Statistical methods,
2. Data-mining methods and

3. Machine learning based methods.

Cluster Analysis for Anomaly Detection

Chandola [6] suggest that, with respect to label availability, anomaly detection can operate in one of three modes:

1. supervised,
2. semi - supervised, and
3. un supervised.

Supervised anomaly detection assumes the availability of a training data set which has in stances labeled as normal or anomalous. Semi - supervised anomaly detection assumes that the training data set includes only normal instances. A model corresponding to normal behavior will be built and used to identify anomalous instances in the test data. Unsupervised anomaly detection does not require any training dataset, instead simply assuming far fewer anomalies than normal instances.

Clustering based techniques for anomaly detection can be grouped into three categories [7]

a. The first group assumes that normal instances belong to a cluster while anomalies do not belong to any cluster. These techniques apply a clustering algorithm to the data set and identify instances that do not belong to a cluster as anomalous.
b. The second group assumes that normal data instances lie closer to the nearest cluster centroid (or center) while anomalies are far away from the nearest cluster centroid
c. The third group assumes that normal data instances belong to large, dense clusters, while anomalies belong to small or sparse clusters.

In [8-11] clustering is established as a useful method for anomaly - based unsupervised detection of intrusions. Clustering techniques have been applied successfully to the anomaly detection problem, where it is applied to the normal samples to generate a set of clusters that will represent the normal class.

Dimensionality Reduction

Most anomaly detection algorithms require a set of purely normal data to train the model and they implicitly assume that anomalies can be treated as patterns not observed before. Since an outlier may be defined as a data point which is very different from the rest of the data, based on some measure, we employ several detection schemes in order to see how efficiently these schemes may deal with the problem of anomaly detection. The statistics community has studied the concept of outliers quite extensively. In these techniques, the data points are modeled using a stochastic distribution and points are determined to be outliers depending upon their relationship with this model. However with increasing dimensionality, it becomes increasingly difficult and inaccurate to estimate the multidimensional distributions of the data points [12].

Statistical and machine reasoning methods face a formidable problem when dealing with such high - dimensional data, and normally the number of input variables is reduced before a data mining algorithm can be successfully applied. The

dimensionality reduction can be made in two different ways: by only keeping the most relevant variables from the original dataset (this technique is called feature selection) or by exploiting the redundancy of the input data and by finding a smaller set of new variables, each being a combination of the input variables, containing basically the same information as the input variables (this technique is called dimensionality reduction) [13].

Principal Component Analysis

Principal Component Analysis is one of the most fundamental tools of dimensionality reduction for extracting effective features from high -dimensional vectors of input data [14,15]. PCA reduces the amount of dimensions required to classify new data and produces a set of principal components, which are orthonormal eigenvalue/eigenvector pairs [16]. It reduces the dimensionality of data by restricting attention to those directions in the feature space in which the variance is greatest. The proportion of the total variance accounted for a feature is proportional to its eigenvalue [17]. PCA is a global transformation and has proven record of high success in reducing dimensions [18]. The main purposes of a principal component analysis are the analysis of data to identify patterns and finding patterns to reduce the dimensions of the dataset with minimal loss of information [19]. In PCA the entire dataset is projected to a different subspace, this answers the disadvantage of the feature selection.

Principal components are a way to picture the structure of the data as completely as possible by using as few variables as possible [20].

For n original variables, n principal components are formed as follows:

- The first principal component is the linear combination of the standardized original variables that has the greatest possible variance.
- Each subsequent principal component is the linear combination of the variables that has the greatest possible variance and is uncorrelated with all previously defined components.

Classification for Anomaly detection

The primary goal of any anomaly detection is to build the identification of the previously unknown system behavior which deviates from normal system behavior. It is a primary task in the anomaly detection is to classify the anomalies and which fall into which category. Anomaly detection is a form of classification [21]. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time. In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on. Credit rating would be the target, the other attributes would be the predictors, and the

data for each customer would constitute a case. Classifications are discrete and do not imply order. Continuous, floating-point values would indicate a numerical, rather than a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm.

The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating. In the model build (training) process, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class assignments are unknown. Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is typically divided into two data sets: one for building the model; the other for testing the model. See "Testing a Classification Model". Scoring a classification model results in class assignments and probabilities for each case.

Classification can be used to detect individual attacks but it has high rate of false alarm. Various algorithms like decision tree induction, Bayesian networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques are used for classification techniques. The classification algorithm has been then applied to audit data collected which then learns to classify new audit data as normal or abnormal data.

Evolutionary Algorithms

Evolutionary approaches have been used by several researchers to optimize intrusion detectors in an automatic manner. Wang et al. used the evolutionary algorithm for discovering neural networks for intrusion detection [22]. Gonzalez et al. proposed an intrusion-detection technique based on evolutionary-generated fuzzy rules [23]. The condition part of the fuzzy detection rules was encoded with binary bits and fitness was evaluated using two factors: the accuracy and the coverage of the rule. The performance was compared to the methods of different genetic algorithms and without the fuzziness of rules.

Genetic algorithms

Genetic algorithms [24] employ metaphor from biology and genetics to iteratively evolve a population of initial individuals to a population of high quality individuals, where each individual represents a solution of the problem to be solved and is compose of a fixed number of genes. The number of possible values of each gene is called the cardinality of the gene. Each individual is called as chromosome. The set of chromosomes forms population.

Existing Methodologies

Cottrell et al. [25] compared different time series forecasting methods to detect anomalies in end to end bandwidth. The data that they used was obtained through active measurements and they were looking for anomalies in available bandwidth over end-to-end links. The paper is valuable because of the comparison of forecasting methods and shows that the forecasting methods used are applicable to several different traffic features. Lazarevic et al. [26] surveyed several clustering based anomaly detection methods and a support vector machine classifier. They evaluated their data on the DARPA 98 dataset and real time internet traffic. The real time traffic was labeled using Snort. The comparison does, however, focus on only one type of anomaly detection method and does not explore the limitations of the methods nor what they are best suited for. Cardenas et al. [27] created a framework for evaluating intrusion detection systems. The paper suggests using a different measure to compensate for the skewed distributions of the problem. We will draw upon their experiences in our evaluation. [28] Compares of different types of anomaly detection methods that detect different types of anomalies.

Problem Formulation

The major problem in the anomaly detection is how to decide the features to be used. The features are usually decided by domain experts. It may be not completely solve the problem. The search for a subset of relevant features introduces an additional layer of complexity in the modeling task. The search in the model hypothesis space is augmented by another dimension, the one of finding the optimal subset of relevant features. It requires more time for learning too. In order to overcome the problems we have to devise the method which could be able solve the disadvantages in the feature selection process.

In this paper we propose a method which involves the dimension reduction technique, which is used to scale down the data for the detection purpose. PCA [16] is a classical method that provides a sequence of best linear approximations to a given high-dimensional observation. It is one of the most popular techniques for dimensionality reduction. The main advantage of this method is easily scale to very high-dimensional datasets, computationally simple and fast, and independent of the classification algorithm. Feature selection needs to be performed only once, and then different classifiers can be evaluated.

But the problems associated with this type of method are they ignore the interaction with the classifier. They are often uni-variate or low-variate. This means that each feature is considered separately, thereby ignoring feature dependencies, which may lead to worse classification performance when compared to other types of feature selection techniques. In order to avoid the problem of lacking of the feature dependencies we cascade the process of dimensionality reduction with the clustering process which follows the dimensionality reduction through PCA.

The multi class classifier is built from the clustered output. This clustering and the multi class classifier are guided by the genetic algorithms for tuning the process to

achieve the better results. The framework of the proposed approach is depicted as per the following figure.

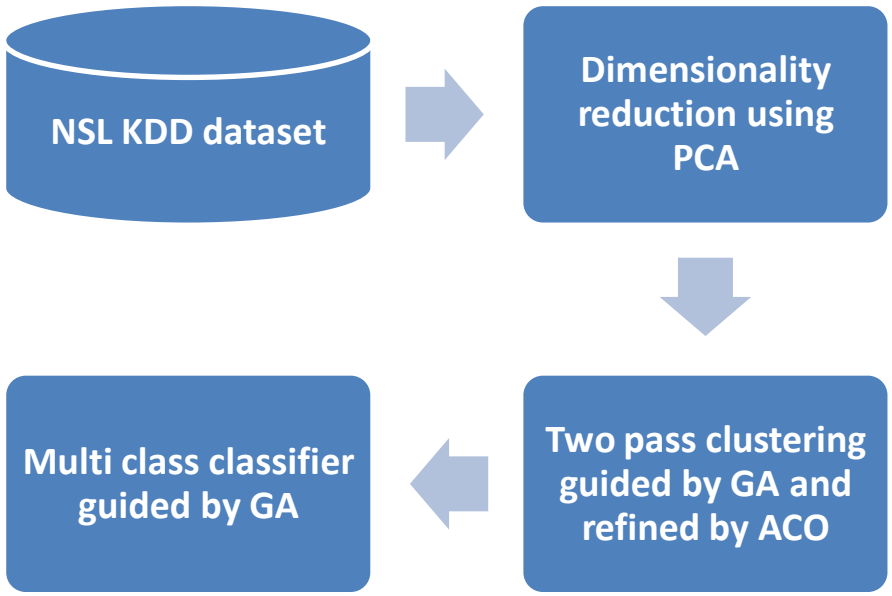

Figure 4.1: Framework of the proposed approach

Proposed Approach
The algorithm of the proposed approach is described as follows.

Input: NSL KDD dataset

Phase I : Dimensionality reduction

Step 1: preprocessing is done by the normalization using Z score
/* a set of n scores each denoted by x_n and whose mean is equal to $\bar{x}$ and whose standard deviation is equal to s is transformed in Z -scores as

$$Z_{n=}\frac{x_n-\bar{x}}{s}$$

*/
Step 2: Application of PCA for the dimensionality reduction
/* PCA algorithm

a. Mean center the data
b. Compute the covariance matrix of the dimensions
c. Find eigenvectors of covariance matrix
d. Sort eigenvectors in decreasing order of Eigen values
e. Project onto eigenvectors in order (The eigenvector with the highest Eigen value is the Principle component of the data)

f. Keep only the terms corresponding to the principle component.

*/

Phase II : Clustering the dimensional reduced dataset

Step 1: /* Here the data is to be clustered by the K means algorithm to find the outlier. The initial cluster centers are determined by genetic algorithm.

a. initialize Y value
b. select Y nodes as initial cluster seed values using genetic algorithm
c. repeat

For J = 1 to n

Compute $|x_i - c_j|^2$ for all cluster seeds

Assign xi to closest cluster cj

Re compute the cluster seed using genetic algorithm

Until (no change in the cluster seed values)

*/

Step 2: /* In this step ACO is used for the refinement of the cluster quality.

a. Input the clusters from pass 1
b. repeat

For i = 1 to n

Use ACO for the adjustments for the clustered items

Repeat

*/

Phase III : Clustering based Multi class classifier guided by genetic algorithm

/* ID3 algorithm to build the decision tree guided by Genetic algorithm

(a) Tree construction

a. choose one attribute as the root with highest information gain and put all its values as branches
b. Apply GA for the choosing recursively internal nodes (attributes) with their proper values as branches for each cluster.
c. Stop when
   - all the samples (records) are of the same class, then the node becomes the leaf labeled with that class
   - or there is no more samples left
   - or there is no more new attributes to be put as the nodes. In this case we apply MAJORITY VOTING to classify the node.

(b) Tree pruning

- Identify and remove branches that reflect noise or outliers using GA based K means Clustering

Output: Multi Class classified dataset

Experiments and Results

The experiment is carried out with the NSL KDD data set. Initially the normalization is done through the Z –score. This method preserve range (maximum and minimum) and introduce the dispersion of the series ie, standard deviation. With elementary algebraic manipulations, it can be shown that a set of Z -score has a mean equal of zero and a standard deviation of one. Therefore, Z -scores constitute a unit free measure which can be used to compare observations measured with different units [29]. PCA is employed for the linear projection of high dimensional data into a lower dimensional subspace. Genetic algorithm based multi class classifier using the decision tree induction is deployed for the classification. The tree pruning is done by the GA based K means clustering.

The results obtained are evaluated based on the following performance metrics

- Entropy
- F measure
- Accuracy
- Error Rate
- Time taken for the classification

The experiment is carried out and the results are evaluated against the classifier proposed in [30]. The results are shown as the graphical representation as follows

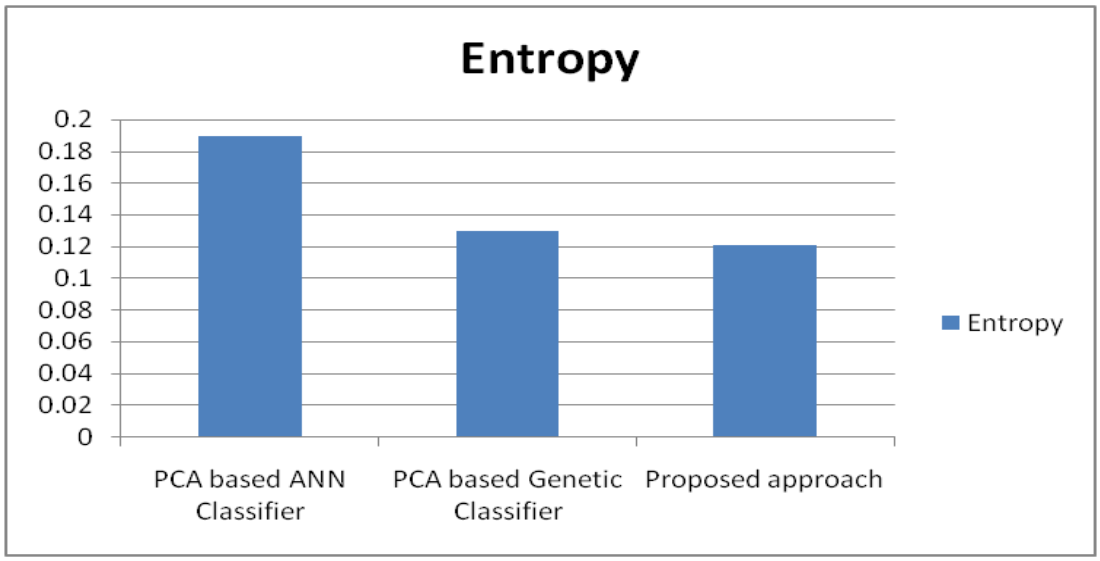

Figure 6.1: Comparison based on entropy measure

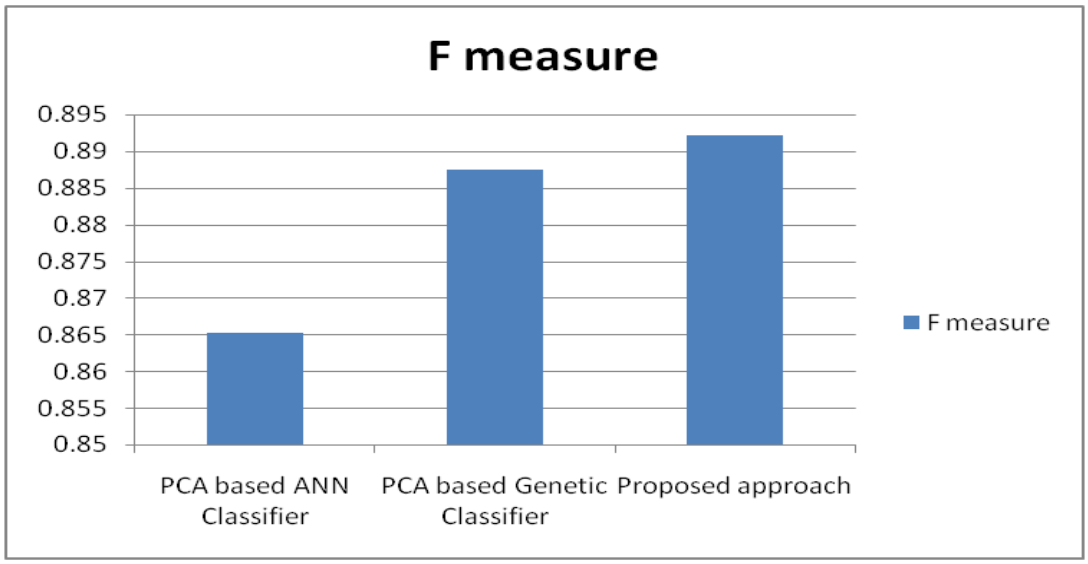

Figure 6.2: Comparison based on F-measure

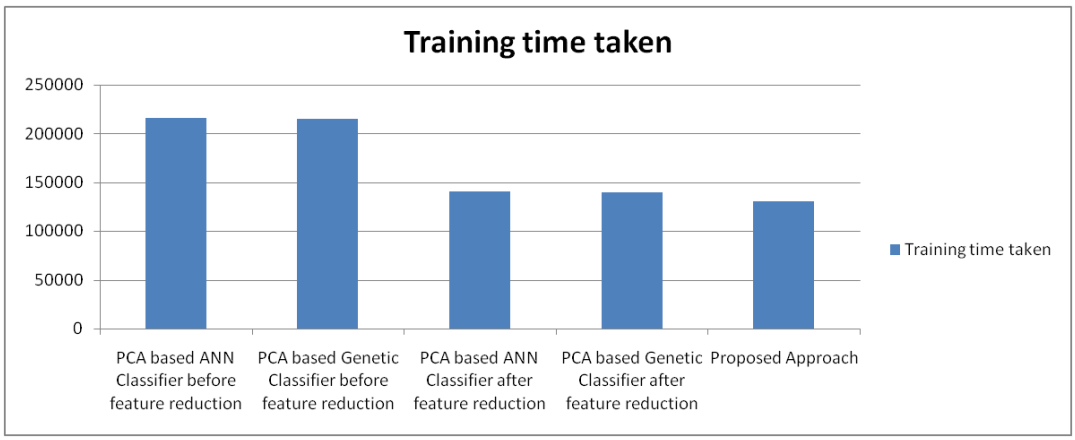

Figure 6.3: Comparison based on Training time Taken

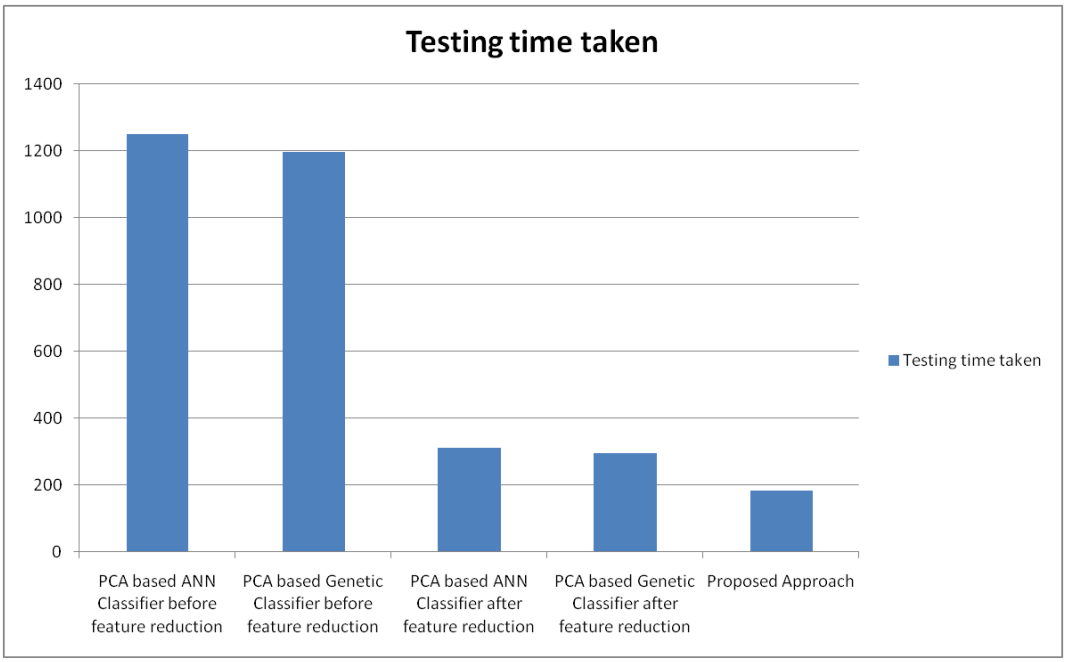

Figure 6.4: Comparison based on testing time taken

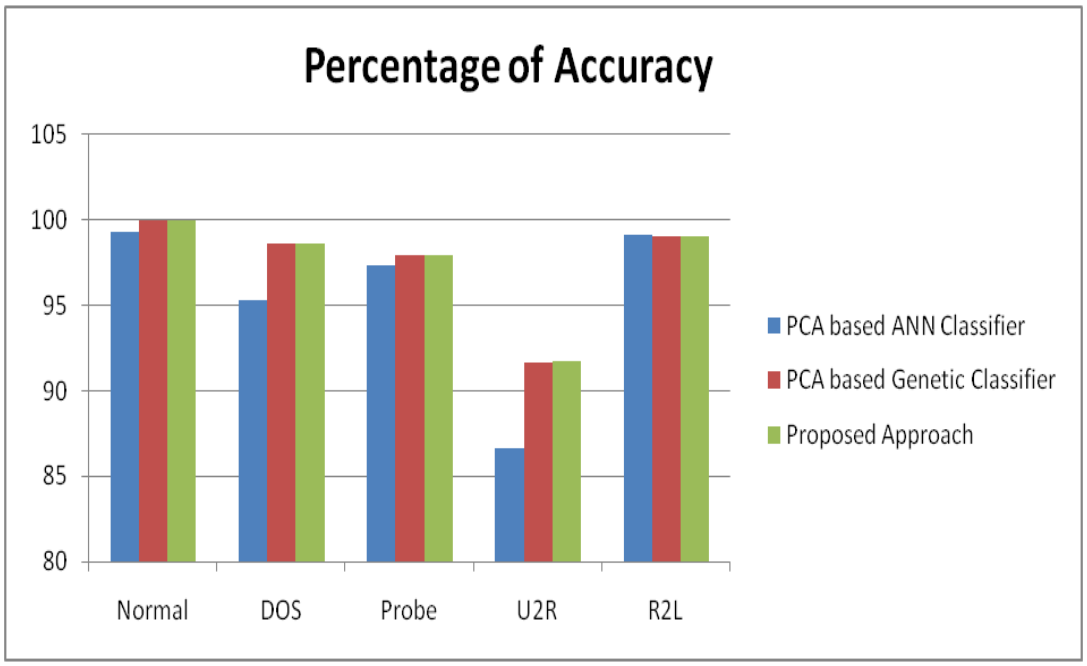

Figure 6.5: Comparison based on prediction accuracy of the classifier

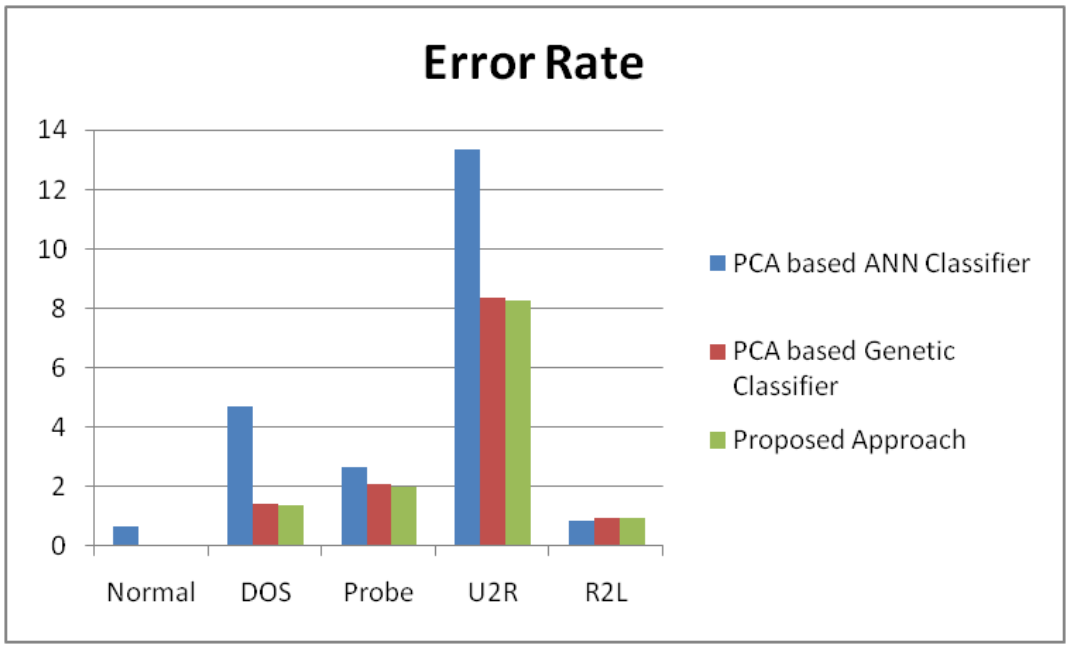

Figure 6.6: Comparison based on error rate of the classifier

Discussion

Table 7.1: Percentage of improvement of the proposed approach based on Entropy

	PCA based ANN Classifier	PCA based Genetic Classifier	Proposed approach	Percentage of Improvement over PCA based Genetic Classifier	Percentage of Improvement over PCA based Genetic Classifier
Entropy	0.1897	0.1293	0.1208	57.03642	7.036424

Table 7.2: Percentage of improvement of the proposed approach based on F-Measure

	PCA based ANN Classifier	PCA based Genetic Classifier	Proposed approach	Percentage of Improvement over PCA based Genetic Classifier	Percentage of Improvement over PCA based Genetic Classifier
F measure	0.8652	0.8874	0.8921	3.031327	0.529637

Conclusion

The risk reduction in the information security is being carried out by the anomaly detection. The classification task is the challenging work in this type of detection. In this paper an optimized approach for the classifier is proposed. PCA is employed for the dimensionality reduction. After the dimensionality reduction clustering process is performed to improve the classification process. Genetic algorithm is employed for the construction of the tree nodes in the building process of the decision tree. The tree pruning is being employed by the clustering approach, which is being guided by the Genetic algorithm. The experimental results are demonstrated and the proposed approach is proven to be the best suited from the classical ANN classifier and the PCA based genetic classifier. The results are promising and the discussions of the experimental results are presented in terms of the percentage of improvement of the proposed approaches over the existing algorithms.

References

[1] Patel A., Qassim Q., and Wills C., "A Survey of Intrusion Detection and Prevention Systems," Information Management and Computer Security, vol. 18, no. 4, pp. 277 290, 2010.

[2] Ujwala Ravale, Nilesh Marathe and Puja Padiya. Article: Attribute Reduction based Hybrid Anomaly Intrusion Detection using K-Means and SVM Classifier. *International Journal of Computer Applications* 82(15):32-35, November 2013

[3] Huebscher M. and Julie A., "A Survey of Autonomic ComputingDegrees, Models, and Applications,"ACM Computing Surveys , vol. 40, no. 3, pp 1-28, 2008.

[4] Qais Qassim, Ahmed Patel, Abdullah Mohd-Zin, Strategy to Reduce False Alarms in Intrusion Detection and Prevention Systems, The International Arab Journal of Information Technology, Vol. 11, No. 5, September 2014

[5] http://www1.cse.wustl.edu/~jain/cse571-07/ftp/ids/index.html

[6] Chandola, V.; banerjee A.; Kumar V.(2009):"Anomaly Detection: A Survey", ACM Computing Surveys , vol. 41, n. 3 :1-58.

[7] Sutapat Thiprungsri, Miklos A. Vasarhelyi, Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach, The International Journal of Digital Accounting Research Vol.11, 2011, pp. 69-84

[8] Leung, K. and Leckie, C. (2005) Unsupervised anomaly detection in network intrusion detection using clusters.Proc. of 28 Australasian conference on Computer Science - Volume 38, Newcastle, NSW, Australia, January/February,pp. 333–342. Australian Computer Society, Inc.Darlinghurst.

[9] Leon, E., Nasraoui, O., and Gomez, J. (2004) Anomaly detection based on unsupervised niche clustering with application to network intrusion detection. IEEE Congres on Evolutionary Computation, 1, 502–508.

[10] Chimphlee, W., Abdullah, A. H., Sap, M. N.M.,Chimphlee, S., and Srinoy, S. (2005) Unsupervised clustering methods for identifying rare events in anomaly detection. Proc. of World Academy of Science, Engineering and Technology, October.

[11] Zhong, S., Khoshgoftaar, T., and Seliya, N.Clustering - based network intrusion detection. Int'nl J of Reliability,Quality and Safety Engineering,14.

[12] K. Hanumantha Rao, G. Srinivas, Ankam Damodhar, M. Vikas Krishna, Implementation of Anomaly Detection Technique Using Machine Learning Algorithms International Journal of Computer Science and Telecommunications [Volume 2, Issue 3, June 2011]

[13] http://arxiv.org/ftp/arxiv/papers/1403/1403.2877.pdf

[14] Lindsay I Smith , " A tutorial on Principal Components Analysis"

[15] CHEN Bo,Ma Wu,Research of Intrusion Detection based on Principal Components Analysis", Information EngineeringInstitute, Dalian University,China,Second International Conference on Information and Computing Science , 2009

[16] I. Jolliffe, Principal Component Analysis . Springer-Verlag, New York, 1986

[17] E. E. Cureton and R. B. D'Agostino, "Factor Analysis: An Applied Approach", London: Lawrence Erlbaum Associates, vol. I, 1983

[18] Veerabhadrappa, Lalitha Rangarajan, Multi-level dimensionality reduction methods using feature selection and feature extraction, International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, October 2010

[19] http://sebastianraschka.com/Articles/2014_pca_step_by_step.html

[20] http://www.jmp.com/support/help/Overview_of_Principal_Component_Analysis.shtml

[21] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/anomalies.htm#CIHGAJAE

[22] L. Wang, G. Yu, G. Wang, and D. Wang, "Method of evolutionary neural network-based intrusion detection," in Proc. Int. Conf. Info-tech and Info-net , Beijing, China, Oct. 2001, vol. 5, pp. 13–18

[23] F. Gonzalez, J. Gomez, M. Kaniganti, and D. Dasgupta, "An evolutionary approach to generate fuzzy anomaly signatures," in Proc. 4th Annu.IEEE Information Assurance Workshop , West Point, NY, Jun. 2003,pp. 251–259

[24] S. Selvakani and R.S. Rajesh, "Genetic Algorithm for Framing Rules for Intrusion Detection" IJCSNS International Journal of Computer Science and Network Security, Vol. 7 No. 11, November 2007

[25] R. Cottrell, C. Logg, M. Chhaparia, M. Grigoriev, F. Haro, F. Nazir, and M. Sandford, "Evaluation of Techniques to Detect Significant Network Performance Problems using End-to-End Active Network Measurements," in IEEE/IFIP Network Operations and Management Symposium NOMS . IEEE, 2006, pp. 85–94

[26] A. Lazarevic, A. Ozgur, L. Ertoz, J. Srivastava, and V. Kumar, "A comparative study of anomaly detection schemes in network intrusion detection," in In Proceedings of the Third SIAM International Conference on Data Mining . Philadel- phia, PA, USA: SIAM, 2003, pp. 25–36.

[27] A. A. Cardenas, J. S. Baras, and K. Seamon, "A Framework for the Evaluation of Intrusion Detection Systems," in Security and Privacy, IEEE Symposium on . Los Alamitos, CA, USA: IEEE Computer Society, 2006, pp. 63–77.

[28] researchcommons.waikato.ac.nz/bitstream/handle/10289/4563/Comparing anomaly detection methods.pdf

[29] Herve Abdi, Normalizing Data, In Neil Salkind (Ed.),Encyclopedia of Research Design.Thousand Oaks, CA: Sage. 2010

[30] Shilpa lakhina, , Sini Joseph, Bhupendra verma, Feature Reduction using Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD, International Journal of Engineering Science and Technology Vol. 2(6), 2010, 1790-1799