Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 7 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Optimized Anomaly based Risk Reduction using PCA based Genetic Classifier

By C. Kavitha & Dr. K. Iyakutti

Pasumpon Muthuramalinga Thevar College, India

Abstract- Security risk analysis is the thrust area for the information based world. The researchers in this field deployed numerous techniques to overcome the information security oriented problem. In this paper the researcher tried for a approach of using anomaly detection for the risk reduction. The hub initiative for this work is that the anomalies are the deviation which could increase the percentage of risk. The anomaly detection is guided by the PCA and the genetic based multi class classifier is used. The classification is induced by the decision tree approach were the genetic algorithm is set out for the optimization in the process of finding the nodes of the tree. The proposed approach is evaluated with the bench mark on PCA based ANN classifier. The proposed approach outperforms the existing one. The results are demonstrated.

Keywords : *anomaly detection, PCA, genetic algorithm.*

GJCST-C Classification : *0*

Strictly as per the compliance and regulations of:

Optimized Anomaly based Risk Reduction using PCA based Genetic Classifier

C. Kavitha [α] & Dr. K. Iyakutti [σ]

Abstract- Security risk analysis is the thrust area for the information based world. The researchers in this field deployed numerous techniques to overcome the information security oriented problem. In this paper the researcher tried for a approach of using anomaly detection for the risk reduction. The hub initiative for this work is that the anomalies are the deviation which could increase the percentage of risk. The anomaly detection is guided by the PCA and the genetic based multi class classifier is used. The classification is induced by the decision tree approach were the genetic algorithm is set out for the optimization in the process of finding the nodes of the tree. The proposed approach is evaluated with the bench mark on PCA based ANN classifier. The proposed approach outperforms the existing one. The results are demonstrated.

Keywords: *anomaly detection, PCA, genetic algorithm.*

I. Introduction

The day to day operations of any enterprise is highly prone to hijack of information. It is very essential part of any enterprise is to put an eye on the security measures so as to minimize the risk of information hijack. The information leaking is not only affecting the day to day activities of the enterprise but also restricts the long term viability of the enterprise. Networks play an important role in the information interchange. Without the communication many of the operation in the gross root level to executive level will be affected. The communication is the back bone of the information and resource sharing. At the same time the network communications are highly riskier because of the intruders. In order to provide a secured information exchange among the enterprise, the first work to be done is to strengthen the network security. The core point of information security is risk management [1]. The people employed with this area use various control and counter measures to defend the security vulnerabilities, but such practices not fulfill the system to be protected against the terrorization due to the intrinsic weakness of the security systems.

Network security measure could be employed on two phases, either the proactive strategy or the reactive strategy. Always the proactive strategy plays a key role for the risk reduction in the information loss and thereby increasing the security level for the enterprise. This paper discuss about a proactive strategy which uses the anomaly detection for the risk reduction. Intrusion detection techniques can be categorized in misuse detection and anomaly detection. Misuse detection systems find intrusions by matching sample data to known intrusive pattern. Anomaly detection systems find intrusion by analyzing the deviation from normal activities profiles that are retrieved from historical data. Intrusion detection is a critical component of secure information systems [2].

The anomaly detection is based on a genetic classifier. This evolutionary strategy is used for the anomaly detection to detect the unusual patterns in the historical data and able to classify the test data. Evolutionary algorithms are a solid but computationally expensive heuristic. Nevertheless, progressively faster computational resources have allowed evolutionary algorithms to be increasingly used in a variety of applications over the years [3]. To facilitate the evolutionary algorithms to come out of the above said drawback Principal Component Analysis (PCA) is employed as a preprocessing measure in this paper.

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.

II. Background Study

This section briefly discuss on the concepts used in this paper. Information security, Anomaly detection, Evolutionary algorithms, Classification, Genetic algorithm, PCA is conferred in this section.

a) *Information security*

In today's fast-changing IT world, even the best available security is insufficient for the latest vulnerabilities in various products, and against malware/attacks created to target those vulnerabilities. While cyber-security cannot be 100 per cent fool-proof, we can still try to achieve the maximum security possible

Author α: Department of Computer Science, Pasumpon Muthuramalinga Thevar College, Madurai, Tamilnadu,India.
e-mail: kkavitha009@gmail.com
Author σ: Department of Physics & Nanotechnology, SRM University,Kattankulathur, Chennai, Tamilnadu, India .
e-mail: iyakutti@gmail.com

[4]. The purpose of the intrusion detection system is to help the computer system on how to deal with the attacks that IDS is collecting information from several different sources within the computer systems and networks and compares this information with pre existing patterns of discrimination as to whether there are attacks or weaknesses [5]. The goal of an intrusion detection system is to provide an indication of a potential or real attack. An attack or intrusion is a transient event, whereas vulnerability represents an exposure, which carries the potential for an attack or intrusion. The difference between an attack and vulnerability, then, is that an attack exists at a particular time, while vulnerability exists independently of the time of observation. Another way to think of this is that an attack is an attempt to exploit vulnerability [6]. An intrusion detection system examines system or network activity to find possible intrusions or attacks. Intrusion detection systems are either network-based or host-based. The information security is highly reliable on intrusion detection in the network since the network is the channel for the intrusion and provides the ways of mishandling the information. The intrusion could be detected in three ways

- Event or Signature-based Analysis
- Statistical Analysis
- Adaptive Systems

The event, or signature-based, systems function much like the anti-virus software with which most people are familiar. The vendor produces a list of patterns that it deems to be suspicious or indicative of an attack; the IDS merely scan the environment looking for a match to the known patterns. The IDS can then respond by taking a user-defined action, sending an alert, or performing additional logging. This is the most common kind of intrusion detection system.

A statistical analysis system builds statistical models of the environment, such as the average length of a telnet session, and then looks for deviations from "normal". After over 10 years of government research, some products are just beginning to incorporate this technology into marketable products.

The adaptive systems start with generalized rules for the environment, then learn, or adapt to, local conditions that would otherwise be unusual. After the initial learning period, the system understands how people interact with the environment, and then warns operators about unusual activities. There is a considerable amount of active research in this area.

Intrusion detection based on anomaly detection techniques has a significant role in protecting networks and systems against harmful activities [7].

b) Anomaly detection

Anomaly Detection is an important alternative detection methodology that has the advantage of defending against new threats not detectable by signature based systems. In general, anomaly detectors build a description of normal activity, by training a model of a system under typical operation, and compare the normal model at run time to detect deviations of interest. Anomaly Detectors may be used over any audit source to both train and test for deviations from the norm [8]. The goal of the anomaly detection is to find all objects that are different to other objects. Anomaly detection finds extensive use in a wide variety of applications such as fraud detection for credit cards, insurance or health care, intrusion detection for cyber-security, fault detection in safety critical systems, and military surveillance for enemy activities [9]. The challenges of the anomaly detection is listed in [9] as follows

- Defining a normal region which encompasses every possible normal behavior is very difficult. In addition, the boundary between normal and anomalous behavior is often not precise. Thus an anomalous observation which lies close to the boundary can actually be normal, and vice-versa.
- When anomalies are the result of malicious actions, the malicious adversaries often adapt themselves to make the anomalous observations appear like normal, thereby making the task of defining normal behavior more difficult.
- In many domains normal behavior keeps evolving and a current notion of normal behavior might not be sufficiently representative in the future.
- The exact notion of an anomaly is different for different application domains. Thus applying a technique developed in one domain to another is not straightforward.
- Availability of labeled data for training/validation of models used by anomaly detection techniques is usually a major issue Often the data contains noise which tends to be similar to the actual anomalies and hence is difficult to distinguish and remove.

c) Evolutionary algorithms

Evolutionary algorithms are stochastic search methods that mimic the metaphor of natural biological evolution [10]. Evolutionary algorithms operate on a population of potential solutions applying the principle of survival of the fittest to produce better and better approximations to a solution. At each generation, a new set of approximations is created by the process of selecting individuals according to their level of fitness in the problem domain and breeding them together using operators borrowed from natural genetics. This process leads to the evolution of populations of individuals that are better suited to their environment than the individuals that they were created from, just as in natural adaptation. Evolutionary computation (EC) techniques can be used in optimization, learning and design [11].

The most important components of any one of the variants of an evolutionary algorithm is given by [12] as

- Representation
- Evaluation function
- Population
- Parent selection mechanism
- Variation operators, recombination and mutation
- Survivor selection mechanism

d) Classification

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute –not yet known. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how "good" the algorithm is [13]. These data analysis help us to provide a better understanding of large data. Classification predicts categorical and prediction models predict continuous valued functions [14].

e) Genetic algorithms

A genetic algorithm is a class of adaptive stochastic optimization algorithms involving search and optimization [15]. Genetic algorithms were first used by Holland (1975). Genetic Algorithms (GAs) are adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. As such they represent an intelligent exploitation of a random search used to solve optimization problems. Although randomized, GAs are by no means random, instead they exploit historical information to direct the search into the region of better performance within the search space. The basic techniques of the GAs are designed to simulate processes in natural systems necessary for evolution; especially those follow the principles first laid down by Charles Darwin of "survival of the fittest." Since in nature, competition among individuals for scanty resources results in the fittest individuals dominating over the weaker ones. [16]

f) Principal Component Analysis

Principal component analysis is appropriate when you have obtained measures on a number of observed variables and wish to develop a smaller number of artificial variables (called principal components) that will account for most of the variance in the observed variables. The principal components may then be used as predictor or criterion variables in subsequent analyses [17]. Principal component analysis (PCA) is a standard tool in modern data analysis - in diverse fields from neuroscience to computer graphics because it is a simple, non-parametric method for extracting relevant information from confusing data sets [18].

III. Existing Methodologies

Security risk assessment and mitigation are two vital processes that need to be executed to maintain a productive IT infrastructure [19]. Risk analysis is the basis of information protection, risk management, and risk in the process of information protection. Risk analysis includes process such as identification of activity, threat analysis, vulnerability analysis and guarantees [20]. The modern security analysis has employed the following techniques like Grey relational making approach [21], Fuzzy number arithmetic operational, Information entropy [22], Fuzzy weighted average approach [23] and Fuzzy measure and Evidence theory [24], Fuzzy AHP method [25-26].

IV. Problem Formulation

In the modern day of communication era Information security highly relies on the network security. In this paper the author tries to employ the anomaly detection mechanism as the base for the identification of the security risk factor. The core idea is the different behavioral pattern leads to the risk. So by the identification of the anomaly could be used to identify the risk.

The major issues in the anomaly detection are the classification mechanism employed. The input data to classifiers is an extremely large set of features, but not all of features are relevant to the classes to be classified. Hence, the learner must generalize from the given examples in order to produce a useful output in new cases. Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features [27].

The advantage of using this method is, it is simple to understand and to interpret. Trees can be visualized. It requires little data preparation. Performs well even if its assumptions are somewhat violated by the true model from which the data were generated. Nonlinear relationships between parameters do not affect tree performance. At the same time we have to concentrate on the issues of Decision tree. Decision-tree learners can create over-complex trees that do not generalize the data well. This is called over fitting. Mechanisms such as pruning, setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.

Decision tree construction with GA Based classification takes longer time then the traditional one although it eliminates many of the unnecessary features; this is one area where more research can be done, to improve the response time [28].

This paper aims to develop a risk reduction mechanism through the employment of the anomaly deduction mechanism through the application of dimension reduction by the PCA and the multi class classifier uses the GA. The pruning is done by the GA based K means clustering.

V. Proposed Approach

The algorithm of the proposed approach is described as follows.

Input: NSL KDD dataset

Step 1: preprocessing is done by the normalization using Z score

/* a set of n scores each denoted by and whose mean is equal to and whose standard deviation is equal to 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.

*/

Step 3: Multi class classifier guided by genetic algorithm

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

Step 3: (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.
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.

Step 3: (b) Tree pruning

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

Output: Multi Class classified dataset

VI. 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
- Feature reduction
- 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

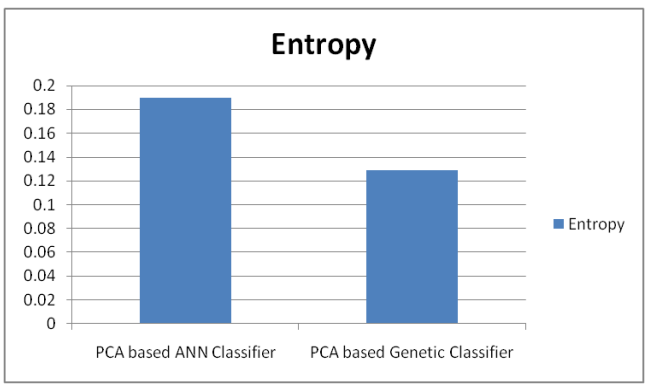

Figure 6.1 : Comparison based on entropy measure

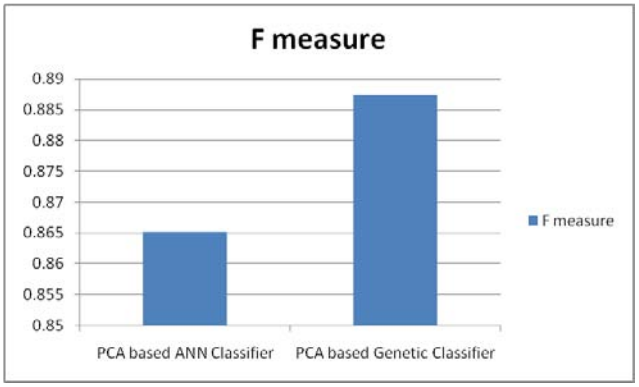

Figure 6.2 : Comparison based on F-measure

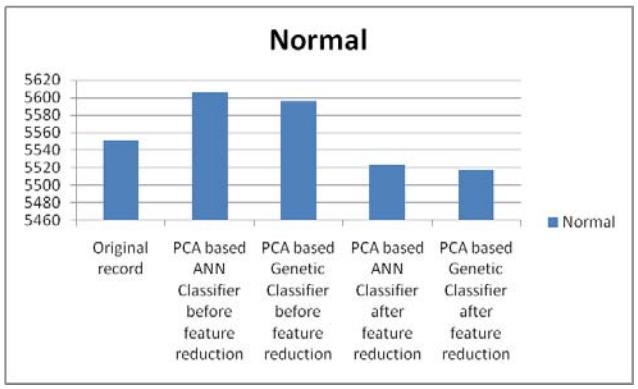

Figure 6.3 : Comparison based on feature reduction on Normal attack data

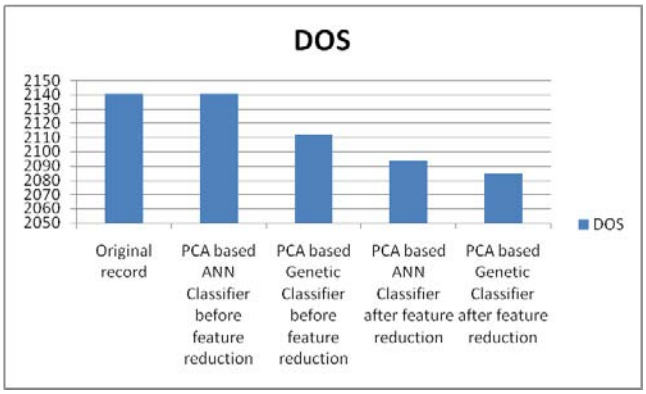

Figure 6.4 : Comparison based on feature reduction on DOS attack data

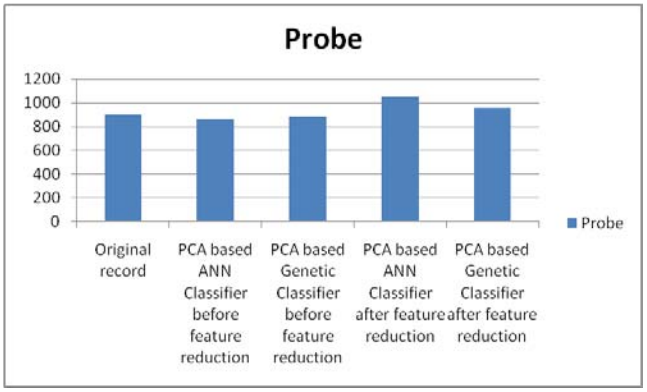

Figure 6.5 : Comparison based on feature reduction on Probe attack data

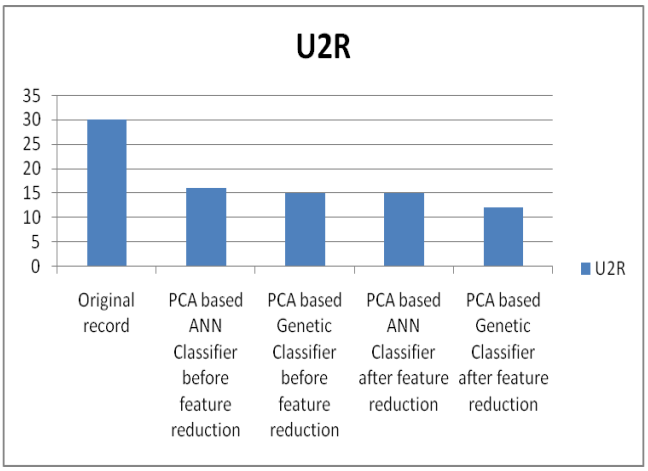

Figure 6.6 : Comparison based on feature reduction on U2R attack data

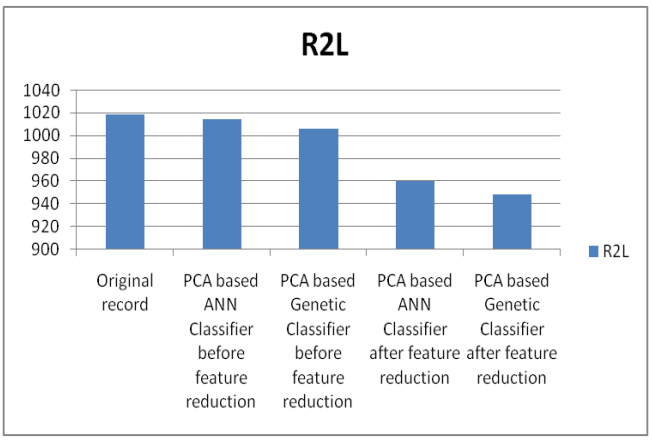

Figure 6.7 : Comparison based on feature reduction on R2L attack data

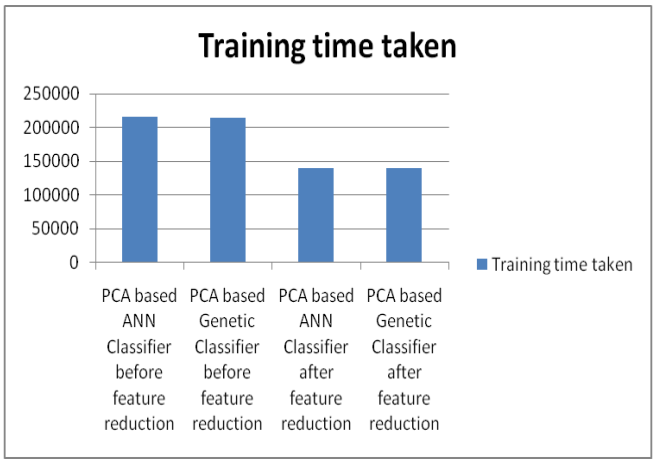

Figure 6.8 : Comparison based on Training time Taken

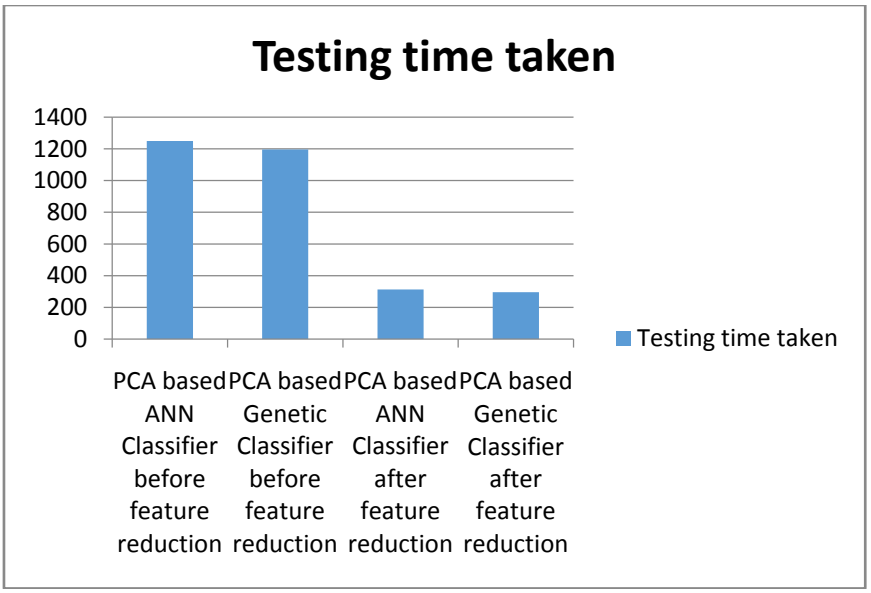

Figure 6.9 : Comparison based on testing time taken

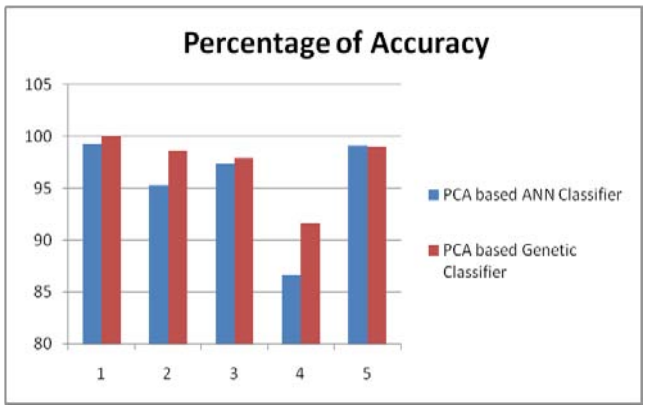

Figure 6.10 : Comparison based on prediction accuracy of the classifier

Figure 6.11 : Comparison based on error rate of the classifier

VII. Discussion

The proposed classifier based on the PCA for the dimensionality reduction and the genetic algorithm with K-means for the decision tree induction is outperforming the existing methodologies.

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

	PCA based ANN Classifier	PCA based Genetic Classifier	Percentage of Improvement
Entropy	0.1897	0.1293	31.83975

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

	PCA based ANN Classifier	PCA based Genetic Classifier	Percentage of Improvement
F measure	0.8652	0.8874	2.50169

VIII. 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. 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.

References Références Referencias

1. Alireza Tamjidyamcholo ,Rawaa Dawoud Al-Dabbagh, Genetic Algorithm Approach for Risk Reduction of Information Security, International Journal of Cyber-Security and Digital Forensics (IJCSDF) 1(1): 59-66, 2012.
2. A. Chandrasekar, V. Vasudevan, P. Yogesh, Evolutionary Approach for Network Anomaly Detection Using Effective Classification, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.1, January 2009.
3. Rodrigo C. Barros, M´arcio P. Basgalupp, Andr´e C. P. L. F. de Carvalho , Alex A. Freitas, A Survey of Evolutionary Algorithms for Decision Tree Induction, IEEE transactions on systems, man, and cybernetics - part c: applications & reviews, Volume:42 , Issue: 3, 2012.
4. http://www.opensourceforu.com/2011/01/importance-of-intrusion-prevention-systems
5. Asmaa Shaker Ashoor, Sharad Gore, Importance of Intrusion Detection System, International Journal of Scientific & Engineering Research, Volume 2, Issue 1, January-2011.
6. https://www.ipa.go.jp/security/fy11/report/contents/intrusion/ids-meeting/idsbg.pdf
7. Tamer F. Ghanema, Wail S. Elkilani, Hatem M. Abdul-kader, A hybrid approach for efficient anomaly detection using metaheuristic methods, Journal of Advanced Research, March 2014.
8. Salvatore J. Stolfo, Shlomo Hershkop, Linh H. Bui, Ryan Ferster, and Ke Wang, Anomaly Detection in Computer Security and an Application to File System Accesses, SMIS 2005, LNAI 3488, pp. 14–28, 2005.
9. Varun Chandola, Arindam Banerjee, and Vipin Kumar, "Anomaly Detection : A Survey", ACM Computing Surveys, Vol. 41(3), Article 15, July 2009.
10. http://www.geatbx.com/docu/algindex01.html#P153_5403
11. http://ewh.ieee.org/cmte/cis/mtsc/ieeecis/Xin_Yao.pdf
12. http://www.cs.vu.nl/~gusz/ecbook/Eiben-Smith Intro2EC-Ch2.pdf
13. http://courses.cs.washington.edu/courses/csep521/07wi/prj/leonardo_fabricio.pdf
14. http://www.tutorialspoint.com/data_mining/dm_classification_prediction.htm
15. http://mathworld.wolfram.com/GeneticAlgorithm.html
16. http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol1/hmw/article1.html#introduction
17. http://support.sas.com/publishing/pubcat/chaps/55129.pdf
18. Jonathon Shlens, A Tutorial on Principal Component Analysis, April 7, 2014; Version 3.02, http://arxiv.org/pdf/1404.1100v1.pdf
19. Nayot Poolsappasit, Rinku Dewri, and Indrajit Ray, Dynamic Security Risk Management Using Bayesian Attack Graphs, IEEE Transaction on Dependable and secure computing, Jan 2012.
20. Ming-Chang Lee, Information Security Risk Analysis Methods and Research Trends: AHP and Fuzzy Comprehensive Method, International Journal of Computer Science & Information Technology (IJCSIT) Vol 6, No1, February 2014.
21. Gao Y, Luo J. Z. (2009), Information security risk assessment based on grey relational decision making algorithm" , Journal of Southeast University , Vol. 39, No. 2, pp. 225-229.
22. Liu F, Dai K, Wang Z. Y. (2004), "Research on the technology of quantitative security evaluation based on fuzzy number arithmetic operation", Fuzzy Systems and Mathematics, Vol. 18, No. 4, pp. 51-54.
23. Chang, P. T., Hung K, C. (2005), "Applying the fuzzy weighted average approach to evaluation network security systems". Computers and Mathematic s with Application, Vol. 49, pp. 1797-1814.
24. Feng N, Li M. (2011), "An information systems security risk assessment model under uncertain environment". Applied Soft Computer, Vol. 11, No.7,pp. 4332-4340.
25. Syamsuddin, I. and Hwang, J., (2010), "The use AHP in security policy decision making: An open office calc application", Journal of Software, Vol. 5, No. 10, pp. 1162-1169.
26. Syamsuddin, I. (2012), "Evaluation of strategic information security with fuzzy AHP method".American Journal of Intelligence Systems, Vol. 2, No. 1, pp. 9-13.
27. http://scikit-learn.org/stable/modules/tree.html
28. Shanta Rangaswamy, Dr. Shobha G, , Sandeep R V, , Raj Kiran, Comparative Study of Decision Tree Classifier with and without GA based feature selection, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 1, January 2014.
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.

Global Journals Inc. (US) Guidelines Handbook 2014

www.GlobalJournals.org