Semantic model for automatic correlation using TLBO for response selection in intrusion detection system

Goldy Saini
Research Scholar, Department of CSE
PIT, Bhopal, India
E-mail-goldysainibe@gmail.com

Deepak Rathore
Asst. Professor
PCST, Bhopal, India
E-mail-rathore.rath@gmail.com

Abstract
In current research trend some authors used some standard technique for feature reduction such as PCA, PCNN and neural network, but these methods not consider all features for processing fixed some number of feature. In this paper we proposed a feature selection and feature reduction method based on improved ID3 algorithm. The proposed algorithm select multiple feature for reduction and the reduce feature set participant the process of detection. The reduce feature of network file classified by ID3 classification algorithm. The ID3 algorithm in the case of small data size, if sizes of data are increase the selection of attribute process raised some problem related to feature selection. For the improvement of this problem used RBF function for increasing the biased value of feature and feature subset selection.

In this paper we tried to propose a very simple and fast feature selection method to eliminate features with no helpful information on them. Result faster learning in process of redundant feature omission. We compared our proposed method with three most successful similarity based feature selection algorithm including Correlation Coefficient, Least Square Regression Error and Maximal Information Compression Index. For the validation and performance evaluation of proposed algorithm used MATLAB software and KDDCUP99 dataset 10%. This dataset contains approx 5 lacks number of instance. The process of result shows that better classification and reduce time instead of another feature reduction.

Keywords: - IDS, KDD CUP, NIDS, TLBO.

1 INTRODUCTION

The performance of intrusion detection system depends on classification of unknown types of attacks. The detection of unknown types of attack is very difficult due to large number of attribute and huge amount of network data. For the improvement of unknown attack feature reduction is important area of research. The reduction process reduces the large number of attribute and improved the detection of intrusion detection system. In the process of feature reduction various algorithm are used such algorithm are principle of component analysis and neural network. The reduction process used PCA method this method is static reduction technique, reduces only fixed number of attribute. The fixed number of feature reduction process not justify the value of feature it directly reduces the feature. On the consideration of computational time feature reduction is also an important aspects, the reduces feature increase the processing of detection ratio. Many methods have been proposed in the last decades on the designs of IDSs based on feature reduction technique. With the tremendous growth of network-based services and sensitive information on networks, network security is becoming more and more importance than ever before. Intrusion detection techniques are the last line of defenses against computer attacks behind secure network architecture design, firewalls, and personal screening. Despite the plethora of intrusion prevention techniques available, attacks against computer systems are still successful. Thus, intrusion detection systems (IDSs) play a vital role in network security. Symantec in a recent report uncovered that the number of fishing attacks targeted at stealing confidential information such as credit card numbers, passwords, and other financial information are on the rise, going from 9 million attacks in June2013 to over 33 millions in less than a year. One solution to this is the use of network intrusion detection systems (NIDS) that detect attacks by observing various network activities. An Intrusion Detection System (IDS) inspects the activities in a system for suspicious behavior or patterns that may indicate system attack or misuse. There are two main categories of intrusion detection techniques; Anomaly detection and Misuse detection. The former analyses the information gathered and compares it to a defined baseline of what is seen as “normal” service behavior, so it has the ability to learn how to detect network attacks that are currently unknown. Misuse Detection is based on signatures for known attacks, so it is only as good as the database of attack signatures that it uses for comparison. Misuse detection has low false positive rate, but cannot detect novel attacks. However, anomaly detection can detect unknown attacks,

but has high false positive rate. There are two primary approaches to analyze events to detect attacks, namely misuse detection and anomaly detection. Misuse detection is based on the extensive knowledge of known attacks and system vulnerabilities provided by a human expert, looking for hackers who attempt to perform these attacks and/or to exploit known vulnerabilities. Although misuse detection can be very accurate in detecting known attacks, it cannot detect unknown and emerging cyber threats this shortcoming makes them vulnerable to the reactivity of attackers. In other words, when attackers change their behavior in response to detection techniques, these techniques become useless and need major redesign. One solution for this problem would be to use adaptive approaches which are inherently designed to be resilient to small changes in the environment and adapt easily. On the other hand, anomaly detection is based on the analysis of profiles that represent normal behavior of users, hosts, or network connections. The rest of paper is organized as follows. In Section II state the problem. The Section III Related work IV discusses proposed methodology. In section V discuss performance evaluation and result analysis followed by a conclusion in Section VI.

II PROBLEM STATEMENT

The environment in which the feature extraction is done is a mobile operator's network with real people (subscribers) using it. This means that the network traffic contains user confidential information. For example in Finland user network traffic is protected by the data protection law. Because of this, only a limited analysis for the network traffic can be done, meaning that a deep packet analysis cannot be done. In general, only the header fields of the packets can be checked but not the user data in the payload. Scalability is an issue with IDS. Because of the huge amount of data flowing through the mobile operator's network, it is not an easy task to find out the right information needed for IDS. The problem is to find an answer to the question: "What features need to be taken into account when calculating or analyzing whether the activity is malicious or not?"Based on prior research on IDS it is clear that either one of the techniques alone cannot detect everything but the combination of the both is the most promising approach. For example misuse detection can be used to filter known threats from the traffic to make it easier for the anomaly detection system to focus on the unknown. Even though IDS have been researched over 20 years, we still do not have an answer to the question of what features should be monitored. So far different kinds of methods and algorithms have been developed for anomaly detection but the focus has been on making them more efficient. Almost all of them are lacking the same information; what features are important for IDS, especially in telecommunications networks? For some reason information on the used features is not easily found from IDS research publications. No matter what the reason is the result is the same; every researcher has to figure out by themselves which features should be used for the monitoring.

1. The pre-processing of KDDCUP99 takes more time.
2. The rate of false alarm generation is high.
3. Some data mining classifier are ambiguous situation for selection of base classifier
4. Entropy based intrusion detection system suffered by high false rate
5. The detection of dynamic feature evaluation as confusion matrix.

III RELATED WORK

We have studied various research and journal papers related to intrusion data classification. According to our research we have analyzed that many of the papers focuses on the problem of better classification of intrusion data and to use an optimized technique for it. Few review of summary described here and implicated with their respective author.

[1] In this paper author tried to propose a very simple and fast feature selection method to eliminate features with no helpful information on them. Result faster learning in process of redundant feature omission. They compared our proposed method with three most successful similarity based feature selection algorithm including Correlation Coefficient, Least Square Regression Error and Maximal Information Compression Index. After that we used recommended features by each of these algorithms in two popular classifiers including: Bayes and KNN classifier to measure the quality of the recommendations. There are varieties of attacks that IDS tries to detect. Some of these can be detected by scanning the packets to find signature of specific attacks.

[2] In this paper author described the feature reduction method with using classifier and the details are Synthetic Minority Oversampling Technique (SMOTE) is applied to the training dataset. A feature selection method based on Information Gain is presented and used to construct a reduced feature subset of NSL-KDD dataset. Random Forests are used as a classifier for the proposed intrusion detection framework. Empirical results show that Random Forests classifier with SMOTE and information gain based feature selection gives better performance in designing IDS that is efficient and effective for network intrusion detection. They used random forests classifier for developing efficient and effective IDS.

[3] In this paper author discussed the techniques for intrusion detection which are based on the data mining techniques and the details are an intrusion detection system (ids) is the fundamental part of the security infrastructure, since it ensures the detection of any suspicious action. Although the detection of intrusions and attacks is the ultimate goal, the huge amount of generated alerts cannot be properly managed by the administrator. In order to improve the accuracy of sensors, we adopt a two-stage technique. The first one aims

to generate meta-alerts through clustering and the second one aims to reduce the rate of false alarms using a binary classification of the generated meta-alerts.

[7] In this paper, a new learning approach for network intrusion detection using naïve Bayesian classifier and ID3 algorithm is presented, which identifies effective attributes from the training dataset, calculates the conditional probabilities for the best attribute values, and then correctly classifies all the examples of training and testing dataset. Most of the current intrusion detection datasets are dynamic, complex and contain large number of attributes. Some of the attributes may be redundant or contribute little for detection making. It has been successfully tested that significant attribute selection is important to design a real world intrusion detection systems (IDS). The purpose of this study is to identify effective attributes from the training dataset to build a classifier for network intrusion detection using data mining algorithms.

[8] In this paper, author use a genetic algorithm to select a subset of input features for decision tree classifiers, with a goal of increasing the detection rate and decreasing the false alarm rate in network intrusion detection. We used the KDDCUP 99 data set to train and test the decision tree classifiers. The experiments show that the resulting decision trees can have better performance than those built with all available features. Machine Learning techniques have recently been extensively applied to intrusion detection. Example approaches include decision trees ,Genetic Algorithm and Genetic Programming, naive Bayes, KNN and neural networks A key problem is how to choose the features (attributes) of the input training data on which learning will take place.

[9] Author proposed here a using SOM for reduce alarm in IDS and described as an Intrusion detection systems aim to identify attacks with a high detection rate and a low false alarm rate. Classification-based data mining models for intrusion detection are often ineffective in dealing with dynamic changes in intrusion patterns and characteristics. Consequently, unsupervised learning methods have been given a closer look for network intrusion detection. Traditional instance-based learning methods can only be used to detect known intrusions, since these methods classify instances based on what they have learned. They rarely detect new intrusions since these intrusion classes has not been able to detect new intrusions as well as known intrusions.

IV PROPOSED METHODOLOGY

Feature correlation and response is very important aspect for intrusion detection and attack prevention. The feature correlation attribute find the correlative feature of attribute of network file. The response factor measures the attack and risk possibility of intrusion detection. In this paper we proposed a feature and response based intrusion data classification. For the extraction of the feature of network file used automatic feature extractor. The automatic feature extractor extracted the feature into different section. And the extracted features pass through the selection process. The process of selection used teacher learning based optimization technique. The teacher learning based optimization technique is basically dynamic population based optimization technique. Here TLBO play a dual role one is feature selection and other is response selection.

The proposed algorithm is combination of feature correlation algorithm and TLBO algorithm. Here feature correlation algorithm generates a input sequence for TLBO algorithm. Basically input sequence is nothing it's a group value of abnormal predicted by our process. Now for the better prediction of risk assessment we optimized a group value of risk correlation. For the optimization of group correlation value we used TLBO algorithm. Here TLBO algorithm minimized a difference of feature attribute.

Input: input a group dataset of data
Output: The feature set of abnormal scenario U kLk, and U k(H (Lk)) (k=1,2,...,n- 1) and input of TLBO algorithm for optimization of feature set.
Algorithm:
1. Find the value of expected feature correlation
Since existence of each abnormal X in a dataset Di is captured by its existence probability P(X, Di)
Feature of an abnormal group S in a dataset Di is the expected of coexistence of all the items in S, i.e.

$$\Pi x \in S\ P(X, Di)$$

The value of expected feature of S in Di is (0, 1]
Feature of an abnormal S in a dataset can be obtained by summing over all datset the expected probability of S, i.e.,

$$\Sigma i\ [\Pi x \in S\ P(x, Di)]$$

2. find all the large 1-abnormal: Lj=tlarge 1-abnormals}; once getting the large (k-l)-abnormal Lk 1, get the large k-abnormal Lk;
3. Given 1i E Lk-l, find the sequence set Pk={P2-P2. I C Lk, where Vp EE Pk 1i is the beginning subsequence of pj; zT(l) = Count(Pk)Icount(li) H(Lk 1) = {zr(li),li E Lk 1};
4. finally, input of TLBO U kLk, and U k(f (Lk)) (k=1,2,...,n-1).
5.
Step 1: Input group of abnormal sequence of data X1,X2…….Xn
Step 2: process in real number and initialize population A(i),i = 0 at random;
Step 3: Calculate the fitness of each individual in the current instant;
FC optimization of expected feature and confidence for generation of feature based tree, Hence the fitness function of algorithm is determined by f(x).

$$F(x) = \begin{cases} (\alpha + 2\beta) - \alpha i, & \alpha i < \beta + 2\alpha \\ 0, & \alpha i \geq \alpha i + 2\beta \end{cases}$$

I=1,2,……..,N

Step 4: Judge the termination conditions. If the termination conditions are satisfied, then turn to step 5, otherwise, turn to step 2;

Step 5: Decode to find and calculate the optimal feature value. And set the feature according to maximum output the results.

V EXPERIMENTAL DETAILS AND RESULT ANALYSIS

In this paper we perform experimental process of proposed Intrusion detection system with using TLBO algorithm. The proposed method implements in mat lab 7.14.0 and tested with very reputed data set from KDD Cup data set. In the research work, I have measured Precision and recall for the S-MAIDS and proposed method. To evaluate these performance parameters we have used KDDCUP99 datasets from UCI machine learning repository.

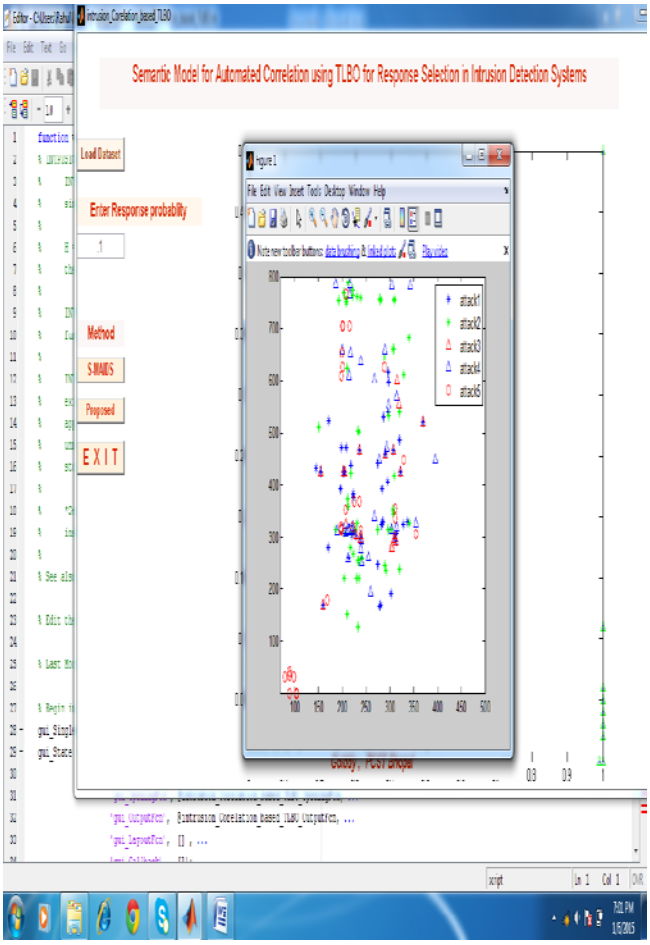

Figure 1: Shows that the result window of attack detection for SMAIDS method using the value is 0.1.

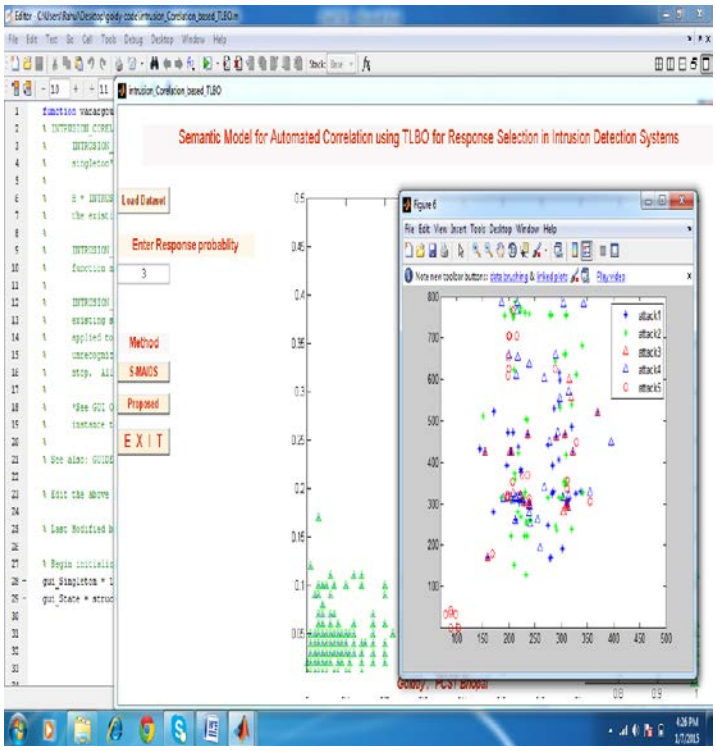

Figure 2: Shows that the result window of attack detection for proposed method using the value is 0.3.

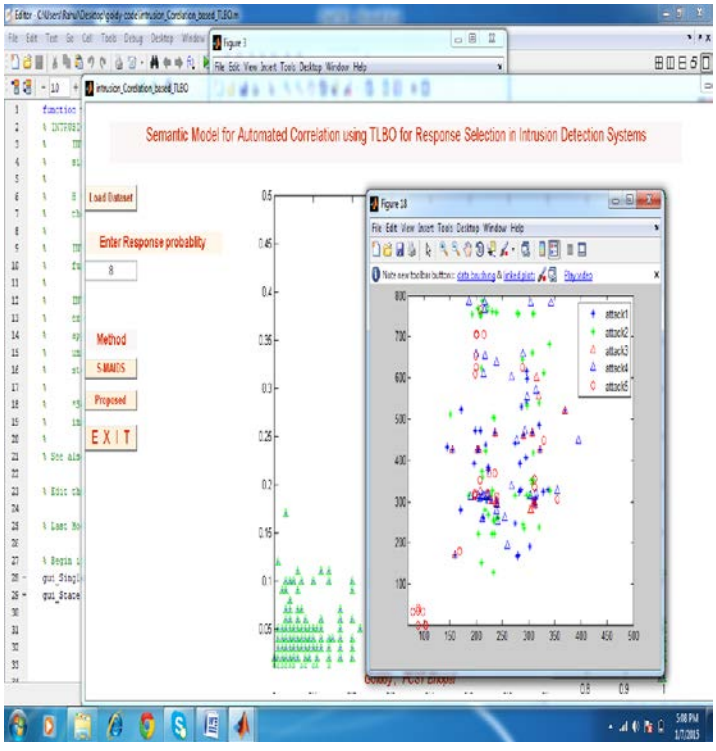

Figure 3: Shows that the result window of attack detection for proposed method using the value is 0.8.

METHOD NAME	NO. OF INPUT	RC	ACR	PCR	NCR
PROPOSED METHOD	.1	95.27	2.85	1.95	2.45
	.2	96.96	4.55	3.65	4.15
	.3	95.27	2.85	1.95	2.45
	.4	97.00	4.59	3.69	4.19
	.5	97.19	4.77	3.87	4.37
	.6	95.27	2.85	1.95	2.45
	.7	97.24	4.83	3.93	4.43
	.8	97.33	4.01	4.01	4.51
	.9	95.27	2.85	1.95	2.45
	.10	95.27	2.85	1.95	2.45

Table 1: Comparative result for intrusion detection with using input values from .1 to .10 for PRPPOSED Method.

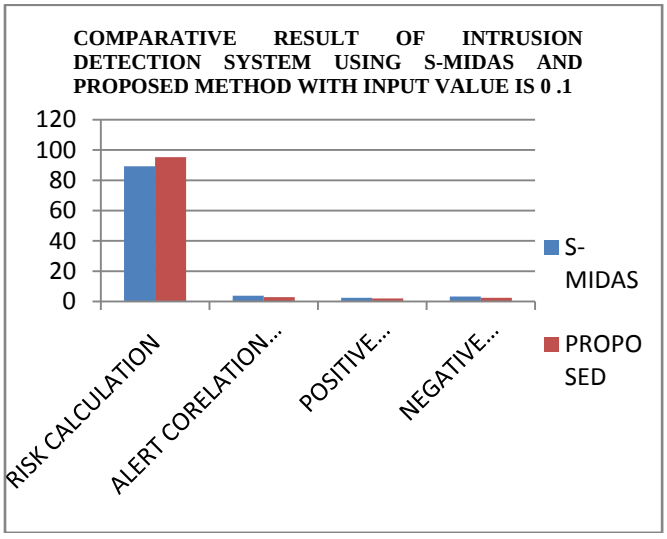

Figure 4: Show that the result graph for Intrusion Detection System using S-MIDAS and proposed method with input value is 0.1, and find the value of risk calculation, alert, positive and negative correlation rate and show that our proposed method gives better result than existing method.

VI CONCLUSION AND FUTURE WORK

In this paper we proposed a feature correlation based response selection of intrusion detection. The feature correlation factor estimates the most similar feature form given dataset. The estimated feature selects the feature value for the selection of response. According to the value of feature similarity select the response of category of abnormal and normal data. In this paper reduction of unwanted attribute of feature selection process is main objective. Because of this consumed time of each algorithm with different reject threshold measured. As evaluation result shows, although FFR cannot defeat other methodologies in accuracy of classification and accuracy didn't changed very much, but in speed FFR outperformed all other feature selection method with great differences. We used FC classifier for developing efficient and effective IDS. For improving the detection rate of the minority classes in imbalanced training dataset we used standard sampling and we picked up all of the important features of the minority class using the minority classes attack mode.

The proposed algorithm is a combination of feature selection and feature response for intrusion detection system. The feature selection and response both improved the performance of classification algorithm, but it not achieved the classification ratio 100%. The process of data sampling improved the reduction process and improved the classification ratio up to 100%. The sampling process design as mixed sampler corresponding to the nature of network traffic data, the network traffic data is mixed data type some are continuous and discrete.

REFERENCES

[1] Shafigh Parsazad, Ehsan Saboori, Amin Allahyar "Fast Feature Reduction in Intrusion Detection Datasets" MIPRO 2012, Pp 1023-1029.

[2] Abebe Tesfahun, D. Lalitha Bhaskari "Intrusion Detection using Random Forests Classifier with SMOTE and Feature Reduction" International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, 2013. Pp 127-132.

[3] Hachmi Fatma, Limam Mohamed "A two-stage technique to improve intrusion detection systems based on data mining algorithms" IEEE, 2013. Pp 1-6.

[4] Shailendra Singh, Sanjay Silakari "An Ensemble Approach for Cyber Attack Detection System: A Generic Framework" 14th ACIS, IEEE 2013.

[5] Li, "Using Genetic Algorithm for Network Intrusion Detection" Proc. the United States Department of Energy Cyber Security Group 2004 Training Conference, May 2004.

[6] Jain , Upendra "An Efficient intrusion detection based on Decision Tree Classifier using feature Reduction", International Journal of scientific and research Publications , Vol. 2, Jan. 2012.

[7] Dewan Md. Farid, Jerome Darmont, Nouria Harbi, Nguyen Huu Hoa, Mohammad Zahidur Rahman "Adaptive Network Intrusion Detection Learning: Attribute Selection and Classification" 2008. Pp 1-5.

[8] Gary Stein, Bing Chen, Annie S. Wu, Kien A. Hua "Decision Tree Classifier For Network Intrusion Detection With GA-based Feature Selection" 2556. Pp 1-6.

[9] Ritu Ranjani Singh a, Prof. Neetesh Gupta "To Reduce the False Alarm in Intrusion Detection System using self Organizing Map" in International journal of Computer Science and its Applications.

[10] Z. Xue-qin, G. Chun-hua, L. Jia-jin "Intrusion detection system based on feature selection and support vector machine" Proc. First International Conference on Communications and Networking in China (ChinaCom '06), Oct. 2006.

[11] Zhang , M. Zulkernine "Network Intrusion Detection using Random Forests" School of Computing Queen's University, Kingston Ontario, 2006.

[12] John Zhong Lei and Ali Ghorbani "Network Intrusion Detection Using an Improved Competitive Learning Neural Network" in Proceedings of the Second Annual Conference on Communication Networks and Services Research IEEE.

[13] P. Jongsuebsuk,N. Wattanapongsakorn and C. Charnsripinyo "Network Intrusion Detection with Fuzzy Genetic Algorithm for Unknown Attacks" in IEEE 2013.

[14] Deepak Rathore and Anurag Jain "A novel method for intrusion detection based on ecc and radial bias feed forward network" in Int. J. of Engg. Sci. & Mgmt. (IJESM), Vol. 2, Issue 3: July-Sep.: 2012.

[15] Wing w. Y. Ng, rocky k. C. Chang and daniel s. Yeung "dimensionality reduction for denial of service detection problems using rbfnn output sensitivity" in Proceedings of the Second International Conference on Machine Learning and Cybernetics, Wan, 2-5 November 2003.

[16] Anshul Chaturvedi and Prof. Vineet Richharia "A Novel Method for Intrusion Detection Based on SARSA and Radial Bias Feed Forward Network (RBFFN)" in international journal of computers & technology vol 7, no 3.

[17] Mohammad Behdad, Luigi Barone, Mohammed Bennamoun and Tim French "Nature-Inspired Techniques in the Context of Fraud Detection" in ieee transactions on systems, man, and cybernetics part c: applications and reviews, vol. 42, no. 6, november 2012.

[18] Alberto Fernandez, Maria Jose del Jesus and Francisco Herrera "On the influence of an adaptive inference system in fuzzy rule based classification system for imbalanced data-sets" in Elsevier Ltd. All rights reserved 2009.

[19] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Macia-Fernandez and E.Vazquez "Anomaly-based network intrusion detection: Techniques, Systems and challenges" in Elsevier Ltd. All rights reserved 2008.

[20] Terrence P. Fries "A Fuzzy-Genetic Approach to Network Intrusion Detection" in GECCO 08, July12–16, 2008, Atlanta, Georgia, USA.