REDUCTION OF DETECTOR GENERATION TIME USING CLUSTERING IN METAHEURISTIC METHOD

[1]Vidhya N. Gavali, [2]Mr. S. M. Sangve
Computer Engineering, Department, ZES's DCOER,Narhe, Pune, India
Email: [1]Vidhya.n.gavali@gmail.com, [2]Sunil.sangve@zealeducation.com

Abstract—**today detection of new threats has become a need for secured communication to provide complete data confidentiality. Network requires anomaly detection to shield from hurtful activities. There are various types of metaheuristic methods have been used for anomaly detection. In this paper, a new approach is described for network anomaly detection by using multi-start metaheuristic method, genetic algorithms and Enhancement in clustering algorithm. A new anomaly detector generation approach based on negative selection algorithm concept is used. Preprocessing on data set, clustering and training dataset selection, detectors generation time and optimization of detectors, rules reduction using Genetic algorithm, training and test dataset evaluation in post processing are the main stages of this approach.**

Keywords- Clustering algorithm, Intrusion detection, Anomaly detection, Negative selection algorithm, Multi-start metaheuristic method, Genetic algorithms

I. INTRODUCTION

The increased network communication in world, security is must for digital information or data. The way of providing security is intrusion detection system (IDS), whose fundamental capacity is to pick up improper, off base and anomalous action in a system. Attacks may be categories into any one of these forms namely Denial of Service (DOS), remote to local (R2L), user to root (U2R) and Probing. In communication networks, intrusion detection may be based on network and/or host or based on the application depending on their mode of deployment and data used for analysis. In the network environment there are two types of intrusion detection systems, anomaly and normal. The normal phase of network is secure, but anomaly network is not secure.

Over the previous decades, a Web and machine system has raised various security issues, because of the unauthorized access of systems and violates system confidentiality, availability and integrity. Any malicious attack on the system may offer to major disaster. Thus, intrusion detection systems (IDSs) are must to diminishing the genuine impact of these attacks [13]. IDSs are named either signature based or anomaly based. Signature based (abuse based) looks for predefined pattern, or signatures. Along these lines, its utilization is best in known attack yet it is unequipped for distinguishing new ones regardless of the possibility that they are fabricated as least variations of known attack. Then again, anomaly-based locators attempt to take in system's typical conduct and create an alert at whatever point a deviation from it happens utilizing a predefined limit value. Anomaly detection can be spoken to as two-class classifier which arranges each one specimen to typical or unusual [15]. It is fit for locating at one time unseen intrusion occasions however with higher false positive rates (FPR, occasions mistakenly named attack) contrasted with signature-based systems [3]. Metaheuristics are

nature based algorithms focused around a few standards from material science, science or ethnology. Metaheuristics are ordered into two fundamental classifications, single arrangement based and populace based Metaheuristics [7]. Populace based Metaheuristics are more suitable in producing irregularity detectors than single-arrangement based because single set of metaheuristic provide single solution. Evolutionary Computation (EC) and Swarm Intelligence (SI) are known types of populace based algorithms. EC algorithms are based on Darwin's evolutionary theory. Genetic algorithms, evolutionary programming, genetic programming, scramble hunt and path re-linking, co-evolutionary algorithms and multi-begin schema [18] are illustrations of EC algorithms. Clustering is a technique where it is possible to find hidden patterns that may exist in datasets and it is possible to infer better conclusions. Clustering techniques are applied in Data Mining and known as "vector quantization" when dealing with Speech and Image data [22]. The most popular Clustering algorithm is K-means (KM) because it can deal with a large amount of data [23], is fast in most cases and it is simple to implement. The main basic idea is to partition the dataset into K clusters. Two weak aspects of KM are the sensitivity to initialization and the combined to local optima [24]. To solve the initialization sensitivity Zhang proposed the K-Harmonic means (KHM).

The Genetic Algorithm (GA) is used in intrusion detection systems (IDSs) to generate rules used to detect anomalies [25].The Genetic algorithms were inspired by the biological evolution (development), natural selection, and genetic recombination. GAs contain three process: selection (called as random selection), cross-over (recombination to produce new data chromosomes), and mutation operators. In last, a fitness function is applied to select the best (highly-fitted) individuals from dataset. The process is done for number of generations until arriving at the individual that nearly meet the sought condition [26], [25]. The genetic algorithm is very efficient in the computer security field, especially in IDSs. GA is usually used to generate rules for intrusion detection, and they usually take the form *if* {*condition*} *then* {*action*}, where the condition part test the fields of incoming network connections to detect the anomalous ones [25].

The negative selection algorithm is a supervised learning algorithm; it is based on the discriminatory mechanism of the natural system. The goal of the negative selection algorithm is to classify a bit or string representations of real-world data as normal or anomalous. It operates in two steps training and testing phases. The basic idea of the negative selection algorithm is to generate a number of detectors in the complementary space and then to apply these detectors to classify new, unseen, data as self pattern or non self pattern. The algorithm can be is summarized in the following steps:

- Define self as a set of S elements of length l over a finite alphabet,
- Then generate a set of D detectors, so that every detector neglects to match any element in S. Rather than definite or prefect matching, the technique utilizes a partial matching rule, in which two strings match if they are identical at least at r contiguous positions, where r is required chosen parameter.
- Monitor S form changes by continually matching the detectors in D against S.

A schematically representation of the algorithm is given in fig. 2. In the original description of the algorithm, candidate detectors are generated randomly and then tested if they match any self string. If a match is found, the candidate is rejected .This process is repeated until a desired number of detectors are generated. A probabilistic analysis is used to estimate the number of detectors that are required to provide a certain level of reliability. The limitation of the random generation approach appears to be computational difficultly to generate valid detectors, which increase exponentially with the size of self (S).

II. LITERATURE REVIEW

Anomaly detection systems work to recognize anomalies in the network environment [13]. At the early stage, the research center using rule-based expert systems to detect anomalies and statistical approaches. But when encountering bigger datasets, the results of rule-based expert systems and statistical approaches had shown less efficiency. Thus, many data mining techniques have been commenced to solve problem among these strategies, the Artificial Neural Network (ANN) is extensively used and

has been successful in solving many complex practical problems [14].The Anomaly network intrusion detection methods are classified into several types. The one of method is Statistic-based method. It identifies the intrusion by using the predefined threshold value, standard deviation, mean, and the probabilities [2], [21]. Another category is Rule-Based methods. It uses the If-Then and If-Else rules. These rules are used to construct the model of detection for some previously known intrusions [12] [19]. Additionally, the State-Based approach is also there. It makes the use of Finite state machine, which is derived from the network topologies to determine the attacks [7] [16]. In addition, heuristic-based approach is also a category for use [1]. Negative selection algorithm (NSA) is one of the artificial immune system (AIS) algorithms which motivated by Immune system microorganism development and tolerance toward oneself in human immune system [6]. The Anomaly detection is attained by building a model of non-typical (non-self) information by producing examples (non-discoverers toward oneself) that don't match existing ordinary (self) designs, then utilizing this model to match non-ordinary examples to recognize peculiarities. In spite of this, models toward oneself (discoverers toward oneself) could be constructed from information toward oneself to identify the deviation from typical conduct [17]. Diverse varieties of NSA have been utilized for anomaly discovery [11]. In spite of the fact that these recently created NSA variations, the fundamental qualities of the first negative selection algorithm [6] still stay, including negative representation of information, circulated area of the detector set which is utilized by matching rules to perform anomaly location focused around separation edge or comparability measure [4]. Producing anomaly detectors required abnormal state arrangement techniques (metaheuristic systems) that give methods to escape from local optima and obtain a solution space for anomaly detection. Multi-start methodology, as one of these systems, were initially considered as an approach to achieve a local or neighborhood seek technique (local solver), by just applying it from various arbitrary starting arrangements. Some kind of expansion is required for looking strategies which are focused around local advancement to cover all arrangement space.

In this paper, the integrated approach is used for anomaly network intrusion detection along with enhancement in clustering algorithm to reduce detector generation time, to increase intrusion detection accuracy and to reduce false positive rate. The Anomaly detection system uses negative selection algorithm, genetic algorithm, and multi-start metaheuristic algorithm together to improve the anomaly detection accuracy.

III. IMPLEMENTATION DETAILS

A. System Overview

Anomaly detection has become an important area of intensive research for secured communication. Many authors have suggested various approaches for unsupervised anomaly intrusion detection with artificial neural networks. The objective of this paper is to classify anomaly or normal class from training dataset. In our architecture having following steps : Training dataset , Pre-processing, Enhanced Clustering algorithms, Training Dataset Selection, Detector Generation and Optimization ,Rule Reduction, Evaluation on Training and Test Dataset ,performance analysis.

In training data set we use NSL-KDD dataset. The dataset contains more records of intrusion pattern using simulated environment to train the model. It consists of 41 attribute in training dataset and contains anomaly and normal class.

The need for data preprocessing can be seen from the fact that redundant data and insignificant features may often confuse the classification algorithm, leading to the discovery of inaccurate or ineffective knowledge. Moreover, the processing time will increase when all features are used. Finally, preprocessing helps to remove the redundant data, incomplete data and transforms the data into a uniform format. The preprocessing module of the proposed system performs the following functionalities:

- Performs redundancy check and handles null values
- Converts categorical data to numerical data

In clustering step we use enhanced clustering algorithm i. e EM Algorithm. The main use of clustering algorithm is to reduce training dataset, increasing processing complexity and reduced time complexity.

In the training data set selection we divide training data set into different cluster and from the each cluster select training dataset sample. Genetic algorithm is used for detector generation .which is focused on the non-overlapping of hyper-sphere detectors to gain the maximal non-self-space coverage by using fitness function which is based on detector radius .In detector generation the normal behaviors of the patterns are called 'self'. This algorithm defines 'self' as normal behavior patterns of a monitored system. It generates a number of random patterns that are compared to each self defined pattern. If any randomly generated pattern matches with the self pattern, this pattern fails to become a detector and thus it is removed. Otherwise, it becomes a 'detector' pattern and monitors subsequent profiled patterns of the monitored system shown in the Fig.1.

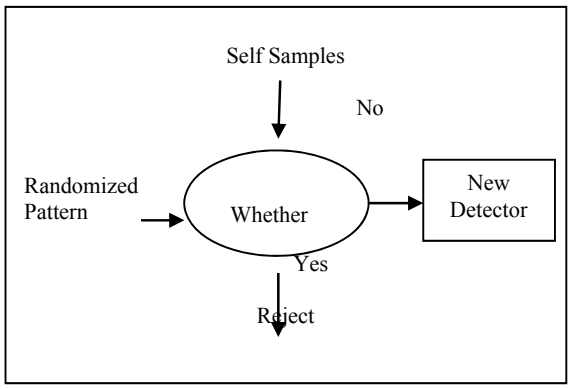

Fig.1. Detector Set Generation using GA [6]

During the monitoring stage, if a 'detector' pattern matches any newly profiled pattern, it is then considered as a new anomaly which must have occurred in the monitored system shown in the Fig.2. The detectors are generated by using GA .All other procedures are similar to negative selection algorithm shown in Fig.2. Initially the KDD Cup 99 dataset is separated as normal file and attack file. From the normal and attack file random data records are chosen for training. The GA parameters are selected. Initial random population is generated and it is evaluated by using fitness function and then the initial population is stored as an initial detector set.

In performance analysis we check the performance algorithm on the basis computation time, accuracy etc.

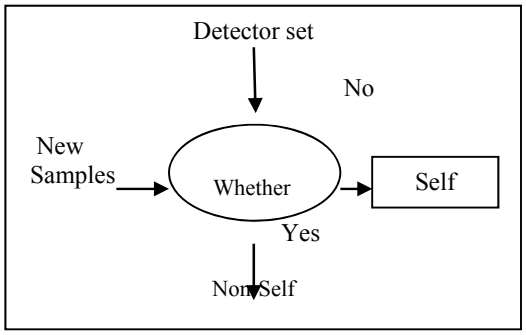

Fig 2 Negative Selection Algorithm [6]

A novel integrated approach for anomaly detection is described in this section. This approach is based on combination of Negative selection algorithm, multi-start metaheuristic algorithm, Genetic algorithm and Enhanced clustering technique. The number of detectors is very important to detect anomaly. In this proposed approach, the main idea is to use enhanced clustering technique to reduce detector generation time and to increase the detection accuracy along with low false positive rate. The enhanced clustering technique is used to select multiple initial points using multi-start method. Using multi-start method, the radius of hyper sphere detector is obtained. This radius is optimized using genetic algorithm. The rule reduction is used to remove redundant detectors to reduce detector generation time. The detector generation process is repeated to increase the detection quality.

The Anomaly detection performance of the system is measured at each step using reduced detectors from earlier step. There are different conditions to stop the repetitive process for example the maximum number of iterations are done without any improvements occurs. At the end the final performance evaluation is based on training dataset which consists of 41 types of attack and testing dataset which consists of additional 14 types of attacks. The anomaly detection is measured using two classifier i.e. normal and abnormal classifier. The system gives detection accuracy, false positive rate and detector generation time.

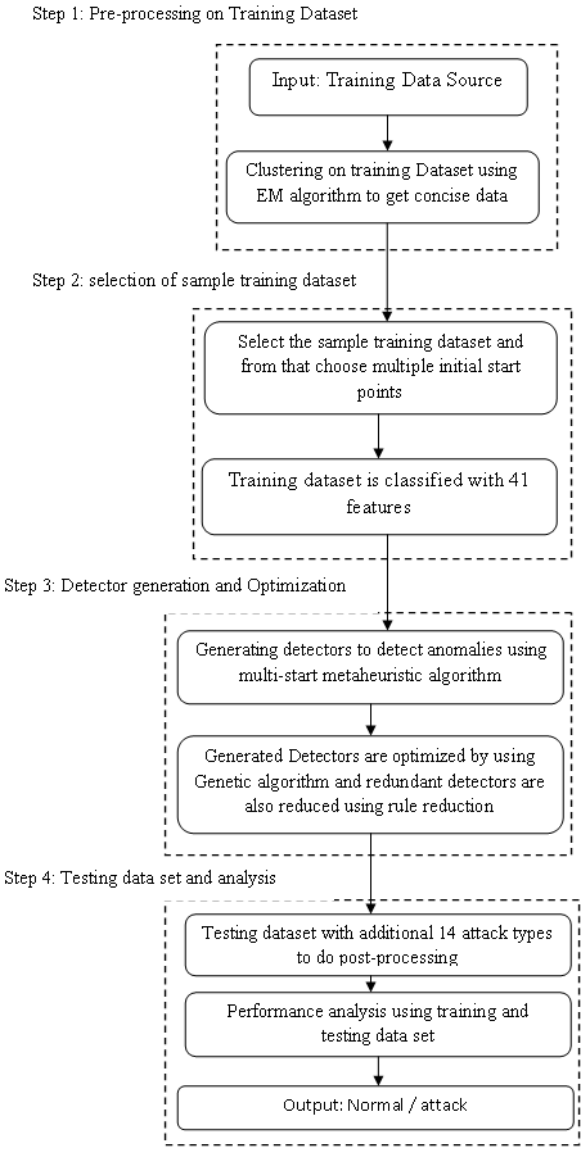

Fig. 3: System Block Diagram

B. Mathematical Model for System

Let, the system S is represented as: S ={ T, P, C, G, R }

A. Training set selection

Let, the T be a training set of trained data

T = {t1; t2; t3; ::; tn }

Where, t1,t2,t3,.. are the number of trained data

B. Preprocessing

Consider, P is a set for processing

D = {p1; p2; p3; :::; pn }

where, p1,p2,p3,... are the number of deterministic variables

C. Clustering

Let, C be a set for k-mediods clustering

C = {c1; c2; c3; ::; cn}

Where, c1,c2,c3,... are the number of clusters

D. Detector Generation and Optimization

Let, G is a set for collecting data

G = {g1; g2; g3; :::; cn}

Where, g1,g2,g3,... are the number of collected data

D. Rule Reduction

Let, R is a set rule reduction

R = {r1; r2; r3; :::; rn}

Where, r1,r2,r3,... are the number of rules in data

Algorithms:

1) Negative Selection algorithm (NSA):

Negative selection algorithm has three stages:

1. Self matching and detection
2. The number of detector generation
3. Monitoring the occurrence of anomalies

Step 1: First, Take the normal pattern. These normal patterns are considered as a „Self" Pattern.

Step 2: Compare the number of random pattern to each self pattern as defined in the self pattern.

Step 3: If any randomly generated pattern matches a self pattern then {this pattern fails to become a detector and thus it is removed} Otherwise {it becomes a „detector"}

Step4: During the monitoring stage, if a detector pattern matches any newly profiled pattern then considered that new anomaly must have occurred in the monitored system**.**

2. Metaheuristic Algorithm (MA):

The Metaheuristic algorithms based on principles of physics and biology. MA has categorized into:

1. One solution approach
2. Population-based metaheuristic algorithms – The examples are co-evolutionary algorithms, genetic programming, scatter search, and multi-start framework.

Step 1: First, perform the robust search for solution space.

This solution space is required for generating detectors to detect anomaly.

Step 2: In this algorithm, the Hyper Sphere detector shape is used to cover most of the normal space.

Step3: To select multi-start parameters initial start points

(ISN) are required. These start points are selected from training data set (TR).

Step4: These selected points are distributed over a normal cluster.

Step 5: There are two solution spaces. One is for upper bound and second solution space for lower bound to cover all the space.

3) EM clustering

In the EM clustering, we use an EM algorithm to find the parameters which maximize the likelihood of the data, assuming that the data is generated from k normal distributions. The algorithm learns both the means and the covariance of the normal distributions. This method requires several inputs which are the data set, the total number of clusters, the maximum error tolerance and the maximum number of iteration.

(1)The EM can be divided into two important steps which are Expectation (E-step)and Maximization (M-step).

(2)The goal of E-step is to calculate the expectation of the
Likelihood (the cluster probabilities) for each instance in the dataset and then re-label the instances based on their probability estimations.

(3)The M-step is used to re-estimate the parameters values from the E-step results.

(4)The outputs of M-step (the parameters values) are then used as inputs for the following E-step.

(5)These two processes are performed iteratively until the results convergence.
The mathematical formulas of EM clustering are described in [15][16] and the pseudo codes can be found in [16].

4. Genetic algorithm:
Genetic algorithms use data which is known as chromosome.
Step 1: selection is initial phase. In selection process random selection is performed.
Step 2: The second step is crossover. The crossover function is used to produce new data using recombination.
Step 3: The third step is to use of mutation operator which is required to match objective function.
Step 4: Finally, Fitness function is applied to select the best individuals. (Fitness function – number of elements in the training set that is covered by the detectors.)
Step 5: The process is repeated for number of generations until reaching desired condition.

IV. RESULTS AND DISCUSSION

A. Dataset

In 1998, DARPA in concert with Lincoln Laboratory at MIT launched the DARPA 1998 dataset for evaluating IDS. The DARPA 1998 dataset contains seven weeks of training and also two weeks of testing data. In total, there are 38 attacks in training data as well as in testing data. The refined version of DARPA dataset which contains only network data (i.e. Tcpdump data) is termed as KDD dataset. The Third International Knowledge Discovery and Data Mining Tools Competition were held in colligation with KDD-99, the Fifth International Conference on Knowledge Discovery and Data Mining. KDD dataset is a dataset employed for this Third International Knowledge Discovery and Data Mining Tools Competition. KDD training dataset consists of relatively 4,900,000 single connection vectors where each single connection vectors consists of 41 features and is marked as either normal or an attack, with exactly one particular attack type. The altered version of KDD Cup 99 dataset is NSL-KDD which gives the better result by reducing redundant data in anomaly network intrusion detection.

B. Experimental Set up

For experimental set up, we use Windows 7 operating system, Intel i5 processor, 512MB RAM, 80GB Hard disk, Net Beans IDE 8 + JDK tool. To calculate the results, NSL KDD data set is used. NSL KDD dataset is a modified version of KDD cup 99. In training dataset, there are 23 types of attack and in testing phase additional 14 attacks are included. Using this dataset, we look for detection accuracy, false positive rate, detector generation time. The detection accuracy is calculated using following formula:

$$\text{Detection Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{(TP+TN+FP+FN)} \quad (1)$$

Where, TP –True Positive
TN-True Negative
FP –False Positive
FN- false Negative

$$\text{False Positive Rate: False Positive} / (FP+TN) \quad (2)$$

Where, True positive means normal sample classified correctly. False positive means normal samples are classified as abnormal incorrectly.

V. CONCLUSION

By using multi-start metaheuristic method with NSA and a genetic algorithm employ a reduction step which is used to remove redundant detectors. It minimizes the number of generated detectors and thus to reduce the time needed later for anomaly detection. The proposed approach with Enhancing in clustering technique is to decrease the detector generation time and also detector radius optimization. There will be Positive effect on processing time if we apply suitable number of detectors with high detection accuracy and low false positive rate. The parameters which are considered as cluster numbers, size of training dataset, staring points, limit of detector radius should be chosen automatically to increase the adaptability and flexibility

REFERENCES

[1] Abadeh MS, Mohamadi H, Habibi J., Design and analysis of genetic fuzzy systems for intrusion detection in computer Networks, Expert Syst Appl 2011.

[2] Assis MVOD, Rodrigues JJPC, Jr. MLP, A hybrid approach for anomaly detection on large-scale networks using HWDS and entropy, in: 21st international conference on software, telecommunications and computer networks (SoftCOM 2013).

[3] Boussaid I, Lepagnot J, Siarry P, A survey on optimization metaheuristics, Inf Sci 2013.

[4] Dasgupta D, Yu S, Nino F, Recent advances in artificial immune systems: Models and applications. Appl Soft Comput 2011.

[5] Esponda F, Forrest S, Helman P, A formal framework for positive and negative detection schemes, IEEE Trans Syst Man Cybernetics, Part B 2004.

[6] Forrest S, Perelson AS, Allen L, Cherukuri R, Self-NonSelf discrimi-nation in a computer, In: Proceedings of the 1994 IEEE symposium on security and privacy; Oakland, USA: IEEE Computer Society; 1994.

[7] Garcia-Teodoro P, Diaz-Verdejo J, Macia - Fernandez G, Va zquez E, Anomalybased network intrusion detection: techniques, systems and challenges. Comput Secur, 2009.

[8] Genetic algorithm. [Online], 2013; ˆahttp://en.wikipedia.org/wiki/Geneticalgorithm ^a:GonazalezF;DasguptaD;valuednegativeselection^a;GenProgramEvolMach; 2003

[9] Gonzalez F, Dasgupta D, Kozma R, Combining negative selection and classification techniques for anomaly detection, In: CEC 02. Proceedings of the 2002 congress on evolutionary computation. Honolulu, HI, USA; 2002.

[10] Ji Z, Dasgupta D, Revisiting negative selection algorithms, Evol Comput 2007.

[11] Kartit A, Saidi A, Bezzazi F, El Marraki M, Radi A, A new approach to intrusion detection system, JATIT 2012.

[12] Liao HJ, Lin CHR, Lin YC, Tung KY, Intrusion detection system: a comprehensive review, J Netw Comput Appl, 2013.

[13] Marti R, Moreno-Vega JM, Duarte A, Advanced multi-start methods, In: Gendreau M, Potvin JY, editors. Handbook of Metaheuristics. Springer US; 2010.

[14] Patcha A, Park JM. An overview of anomaly detection techniques: existing solutions and latest technological trends. Comput Netw, 2007.

[15] Shameli Sendi A, Dagenais M, Jabbarifar M, Couture M, Real time intrusion prediction based on optimized alerts with hidden Markov model, JNW 2012.

[16] Stibor T, Mohr P, Timmis J, Eckert C, Is negative selection appropriate for anomaly detection, In: Beyer HG, editor. GECCO ˆa05 Proceedings of the 2005 conf. on genetic and evolutionary computation. New York, NY, USA: ACM, 2005.

[17] Ugray Z, Lasdon L, Plummer J, Glover F, Kelly J, Marti R, Scatter search and local NLP solvers: a multi-start framework for global optimization, Informs J Comput 2007.

[18] Wang SS, Yan KQ, Wang SC, Liu CW, An integrated intrusion detection system for cluster-based wireless sensor networks, Exper Syst Appl 2011.

[19] [20].Wu SX, BanzhafW, The use of computational intelligence in intrusion detection systems: a review, Appl Soft Comput 2010.

[20] Xu X, Sequential anomaly detection based on temporal difference learning: principles, models and case studies. Appl Soft Comput 2010.

[21]Z. G¨ung¨or and A. Unler, "K-harmonic means data clus- ¨ tering with simulated annealing heuristic." Applied Mathematics and Computation, vol. 184, no. 2, pp. 199–209, 2007.

[22] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. USA: University of California Press, 1967, pp. 281–297.

[23] S. Z. Selim and M. A. Ismail, "K–means type algorithms: A generalized convergence theorem and characterization of local optimality," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 6, pp. 81–86, 1984.

[24] B. Zhang, M. Hsu, and U. Dayal, "K-harmonic means - a data clustering algorithm," Hewlett-Packard Laboratories, Palo Alto, Tech. Rep. HPL-1999-124, Outubro 1999.

[25] Wei Li, "Using Genetic Algorithm for Network Intrusion Detection",Proceedings of the United States Department of Energy Cyber Security Grou, Training Conference, Vol. 8, pp. 24-27,2004.

[26] Fabio A. Gonzalez and Dipankar Dasgupta, "An Immunity-based Techniqueto Characterize Intrusions in Computer Networks".

[27] Chen, Q. and Aickelin, U. Dempster–Shafer for anomaly detection, Proceedings of the International Conference on Data Mining DMIN 2006, Las Vegas, USA, pp. 232–238 (2006).

[28] Wang, Gang, Hao, Jinxing, Ma, Jian and Huang, Lihua „„A new approach to intrusion detection using artificial neural networks and fuzzy clustering"",Original Research Article Expert Systems with Applications, 37(9), pp. 6225–6232 (2010).

[29]Seetha, J., Varadharajan,R., Vaithiyananthan, V. Unsupervised Learning Algorithm for Color Texture Segmentation Based Multiscale Image Fusion. European Journal of Scientific Research, ISSN 1450-216X, Vol 67, pp. 506-511 (2012)

[30]. Lu, Wei., Tong, Hengjian.: Detecting Network Anomalies Using CUSUM and EM Clustering. Advance in Computation and Intelligence, Volume 5821, pp. 297-308 (2009)