International Journal of Security and Its Applications
Vol.8, No.5 (2014), pp.109-118
http://dx.doi.org/10.14257/ijsia.2014.8.5.11

Alert Management System using K-means Based Genetic for IDS

Mohammad Masdari[1] and Fatemeh Charlank Bakhtiari[2]

[1]Department of Computer Engineering, Islamic Azad University, Science and Research Branch, Urmia
[2]Department of Computer Engineering, Islamic Azad University, Science and Research Branch, Urmia
[1]m.masdari@iaurmia.ac.ir, [2]Fatemeh.charlangbakhtiari@yahoo.com

Abstract

One of the most important tools in security field is Intrusion Detection System. The aim of the IDS is to monitor suspicious network traffic and generate alerts. These systems are known to generate numerousfalse positive alerts. Analyzing the alerts manually by security expert need more time and could be error prone.Another problem with IDS is Identifying attack types and generating correct alerts related to attacks.we introducenew alert management systems to overcome mentioned problems. Alert management systems help security experts to manage alerts and produce a high level view of alerts.

In this paper a new alert clustering algorithm for IDS Alert Management System proposed that uses the K-mean Based Genetic (KBG). The proposed algorithm reduces alerts and detects false positive alerts. By the experimental results on DARPA KDD cup 98 the system is able to cluster and classify alerts and causes reducing false positive alerts considerably.

Keywords: *IDS, Alert management, Artificial Neural Network, false positive alert reduction*

1. Introduction

An Intrusion Detection System (IDS) is a hardware device or software program that analyzes computer system activities and/or network traffics to detect malicious activities and produces alerts to security experts [1]. These systems generate lots of false positive alerts. Manually analyzing these alerts by security expert as a problem of ids, need more time and could be error prone [2]. Another problem with IDS is Identifying attack types and generating correct alerts related to attacks. Our proposed system is to overcome mentioned problems. Alert management systems help security experts to manage alerts and produce a high level view of alerts.

Now we describe types of alerts as a need for this article.

False Positive, False Negative, True Positive, True Negative

False positive problem is mystery term that means the IDS generated alerts for one malicious activity but it generate alerts for normal activity (IDS makes a mistake) [27, 28]. Organizing and dealing with the recorded logs and generated alerts by the security sensors such as the IDS, firewalls, packet filtering and servers are not easy job. Most of the organizations consider these alerts as a major problem. Since these sensors are independent so they will generate alerts and send it to the analyst part. They analyze these alerts for understanding the nature of the intrusion, using the provided tools, methods and techniques leading increase the attack detection rate and to reduce the false alerts rate. Even after, huge

ISSN: 1738-9976 IJSIA

number of alerts with a plenty of false alerts will be the way of how any sensor works even when a harmless event accrued.

When an attack has taken place and no alarm is raised known it is the false negatives. False negative can also be defined as an action of IDS system that does not detect actual anomaly/misuse action and allows passing. Subject's normal behavior is the basis for the Anomaly detection, "any action that significantly deviates from the normal behavior is considered as intrusive". Therefore the normal behavior in IDS shall be defined explicitly. Stefano Zanero [29] proposed models for the evaluation of the IDS. Anomaly detection systems report more false positives and less false negatives in while; signature based systems report very false positives, but produces more false negatives. J Snyder [30] states that "the target-based architectures will reduce false positives". False negatives also create a nuisance and issue of importance. Large number of new attacks will generate false negatives in misuse based systems, since there may not be any similar signature.

True positive alerts mean: A legitimate attack which triggers an IDS to produce an alarm. [31]

True negative alerts mean: An event when no attack has taken place and no detection is made.

According to the real nature of a given event and the prediction from IDS, four possible outcomes are shown in Table I, which is known as the confusion matrix [32, 33]. True negatives as well as true positives correspond to a correct operation of the IDS; True negatives (TN) are events which are actually normal and are successfully labeled as normal, true positives (TP) are events which are actually attacks and are successfully labeled as attacks. Respectively, false positives (FP) refer to normal events being classified as attacks; false negatives (FN) are attack events incorrectly classified as normal events.

False positive rate (FPR) also known as false alarm rate (FAR), defined as the number of normal patterns classified as attacks (False Positive) divided by the total number of normal patterns. A high FPR will seriously cause the low performance of the IDS and a high FNR will leave the system vulnerable to intrusions. TNR also known as detection rate or sensitivity refers to proportion of detected attacks among all attack events. Accuracy refers to the proportion of events classified as an accurate type in total events [33]. So, to have effective IDS both FP and FN rates should be minimized, also TP and TN rates must be maximized too.

Nowadays, intrusion detection system requires high detection rate and low false alarm rate. The important issue about evaluating different algorithms is reducing false positive rate but is not enough? Some false positive reduction techniques will cause low accuracy because of some operations like over generalization, missing real attack alerts, *etc.*, So, effective techniques will reduce the false positives rates while increase the accuracy of the system or at least keep it without change.

One of the methods of alert management is clustering of alerts. According to the recent researches, clustering of alerts is an NP-Complete problem [8].In this paper by using K-mean Based Genetic (KBG), an alert management system is proposed which classifies alerts and detects false positive alerts. The system uses generated clusters as a classifier to identify new alerts. This paper proposes a new technique to use generated cluster from GA clustering as a classifier. To improve accuracy of the results, the proposed system uses some techniques such as alert filtering and alert preprocessing. The alert management system is introduced in Section 1. Section 2 reviews related works, Section 3 explains the suggested alert management system. The experimental results are shown in Section 4 and finally Section 5 is conclusion and future works.

2. Related Works

Clustering of alerts is an example of alert management techniques. In [2] a clustering method is introduced based on discovering root cause of false positive alerts. Removing the root causes enhances alarms quality in the future. The root cause instigates the IDS to trigger alarms that almost always have similar features. These similar alarms can be clustered together; consequently, we have designed a new clustering technique to group IDS alarms and to produce clusters. The results in [2] show that a small number of root causes implies 90% of alerts. By removing these root causes total number of alerts come down to 18%. One of the main problems of this technique is depending on under laying network structure.

In [16] some heuristic and neural network based techniques are used to cluster alerts. The main contributions of [16] are: 1) the use and analyses of real network data (data recorded from an existing critical infrastructure); 2) the development of a specific window based feature extraction technique; 3) the construction of training dataset using randomly generated intrusion vectors; 4) the use of a combination of two neural network learning algorithms - the error-back propagation and Levenberg-Marquardt, for normal behavior modeling. Another clustering technique is used in Mirador project with expert systems by Cuppens that similarity between two alerts is calculated by expert system [3, 4]. The results of two genetic clustering algorithms, named Genetic Algorithm (GA) and Immune based Genetic Algorithm (IGA) are compared in [5, 6]. After analyzing the characteristics of Immunity Intrusion Detection System, by utilizing prominent characteristics of genetic algorithm and vaccine mechanism, a new hybird immunity intrusion detection model based on genetic algorithm and vaccine mechanism was established.

Wespi and Debar [16] design an algorithm that places alerts in situations. Situations are set of special alerts and are created with source, destination and attack class attributes.

In [17] an algorithm is introduced that create hyper alert from existing alerts. A hyper alert is an aggregation of related alerts. An alert management system is introduced in [7] that used Self-Organizing Maps (SOM) to cluster and classify IDS alerts. In [7] several operations and methods such as alert filtering, alert preprocessing and cluster merging are introduced.

In this paper an alert management system similar to system [8] are designed that uses K-means Based Genetic (KBG) to classify generated alerts. The system will be able to improve accuracy of results, to identify attack type of alerts accurately and also to reduce the number of false positive alerts considerably.

3. Proposed Alert Management System Based on Kbg

Figure 1 shows the proposed system. In this paper we use binary traffics files of a network, DARPA 98 dataset [10] instead of real network traffics. Snort tool [11] is used to produce alerts of DARPA 98 dataset network traffics. Snort is an open source signature based IDS which gets DARPA 98 online traffic and then generates alert log files [8]. After generating alert log files, these files are entered to the proposed system as the inputs.

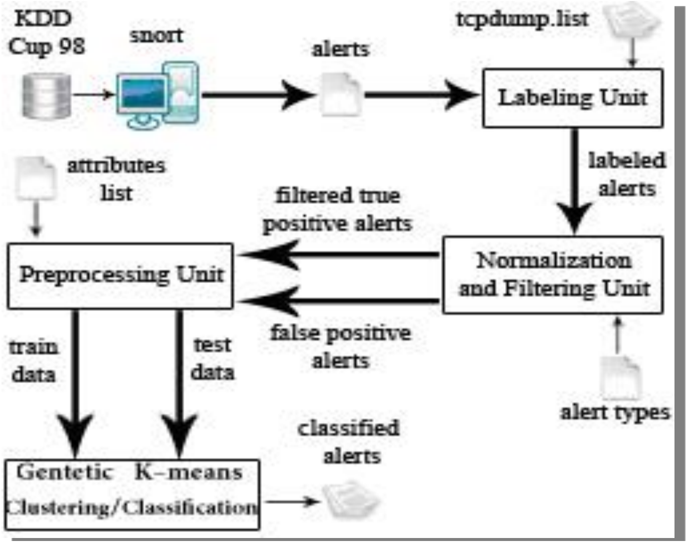

Figure 1. Proposed Alert Management System

3.1. Labeling Unit

Labeling unit gets generated alert from Snort and also tcpdump. List files of DARPA 98 dataset and then generate labeled alerts. A labeled alert is an alert with its own attack type. The tcpdump. list files contain information about all packets in DARPA 98 dataset. These labels are used to train KBG and evaluate results of KBG [8, 9].

3.2. Normalization and Filtering Unit

In this phase accepted attack types are entered to the unit and only alertsare selected that are in class of predefined attack types [8, 9 and 12]. This unit uses eight attributes of alert to filter alert, which are: Signature ID, Signature Rev, Source IP, Destination IP, Source Port, Destination Port, Datagram length and Protocol [12].

3.3. Preprocessing Unit

Preprocessing unit converts string values of attributes of alert to numerical data. It also reduces the range of attribute values and converts alerts to data vectors (1), (2) and (3).

$$IP = X_1 . X_2 . X_3 . X_4 ,$$
$$IP\_VAL = (((X_1 \times 255) + X_2) \times 255 + X_3) \times 255 + X_4 \qquad (1)$$

$$protocol\_val = \begin{cases} 0, \; protocol = None \\ 4, \; protocol = ICMP \\ 10, \; protocol = TCP \\ 17, \; protocol = UDP \end{cases} \qquad (2)$$

$$IUR = 0.8 \times \frac{x - x_{\min}}{x_{\max} - x_{\min}} + 0.1 \qquad (3)$$

3.4. K-means Based Genetic Algorithm (Cluster/Classify) Unit

1) Genetic Algorithm

Genetic Algorithms (GAs) are optimization and search procedures which uses the principles of natural selection in the natural genetics [18-21].

In GA, the role of crossover, mutation and selection operators is well defined. The direction of the search is controlled with selection operator. The crossover operator constructs new search region. The role of mutation operator is exploring the search space. "GAs perform search in complex, large and multimodal landscapes and provide near optimal solutions for objective or fitness function of an optimization problem" [23].

In GAs, the parameters of search space or optimization problem are encoded into strings named chromosomes that collection of such chromosomes is called a population. In the initialization phase, the population is created randomly. Each chromosome in population presents a point in the search space. A fitness or objective function is associated with each chromosome in population that represents the degree of goodness of the proper chromosome. Based on the survival of the fittest, a few of the chromosomes are selected and is assigned a number of copies to each gathering into go into the mating pool. To constructing new generation from this population, crossover and mutation operators are applied. The sequence of selection, crossover and mutation operators continue for either a fixed number of generations or termination condition is satisfied [19].

2) Overview of GA based Clustering Algorithms

Cluster analysis is one of the techniques to discover patterns and associations within data. "A clustering method is a multivariate statistical procedure that starts with a data set containing information about a sample of entities and attempts to reorganize these entities into relatively homogeneous groups" [24]. One of the main problems encountered by researches, with regard to cluster analysis is that different clustering techniques can construct different solution for the same set of data. In this case a technique is needed to discover the most natural groups in the set of data.

The research effort by Krovi R. was to investigate the potential feasibility of using genetic algorithms for the purpose of clustering [23].

The encoding of a chromosome results a string of real numbers. A chromosome represents a solution that contains centroids of generated clusters. Each chromosome contains sets of genes (the number of genes in each set equal to number of attributes in search space) corresponding to a cluster centroid.

The fitness calculation process consists of two phases. In the first phase, the clusters are formed according to the centres encoded in the chromosome under consideration. This is done by assigning each point x_i, i=1, 2,.., n, to one of the clusters C_j with centre z_j such as [24]

$$\left\|x_i - z_j\right\| < \left\|x_i - z_p\right\|, p = 1,2,..., k, and \ (p \neq j) \quad (4)$$

Then the new centroids are calculated according to

$$z_i^* = \frac{1}{n_i} \sum_{x_j \in C_i} x_j, i = 1,2,..., K \quad (5)$$

wherez_i^* is the new centroid and n_i is the number of points in the cluster i. After calculating new cluster centroids, cluster metrics must be computed for each cluster. It is the sum of the Euclidean distances of the points from their proper cluster centres (6) [24].

$$M(C_1, C_2, ..., C_K) = \sum_{i=1}^{K} \sum_{x_j \in C_i} \left\|x_j - z_i\right\| \quad (6)$$

whereC_i is the cluster, z_i is the centre of the cluster C_i and x_j is a point in cluster C_i.

The fitness function for the GA is shown in (7) [24].

$$M = \sum_{i=1}^{K} M_i$$

$$M_i = \sum_{x_j \in C_i} \left\| x_j - z_i \right\|.$$

$$f = 1 / M \qquad (7)$$

This paper uses roulette wheel for selection operator and two point crossover for crossover operation.

Since the paper uses floating point representation the following mutation is used. A value ∂ in unit range ([0, 1]) is generated with uniform distribution. If the value at a gene position is v after calculating by mutation [24].

$$v \pm \partial * v \qquad (8)$$

After the fix number of iterations, the termination function decides to stop the running. So the termination condition is the number of iterations.

The four parameters (Fitness Function, Selection Function, Mutation Function and Termination Function) are similar in the other entire GA-based clustering algorithm evaluated in this paper.

4. KBG Clustering

In KBG, GA uses an extra operator named K-Means [25] Operator (KMO) .The GA uses the mutation and the selection operators as mentioned before, but the algorithm with these operators may take more time to converge. To solve this problem, the algorithm uses one step K-means algorithm instead of crossover operation.

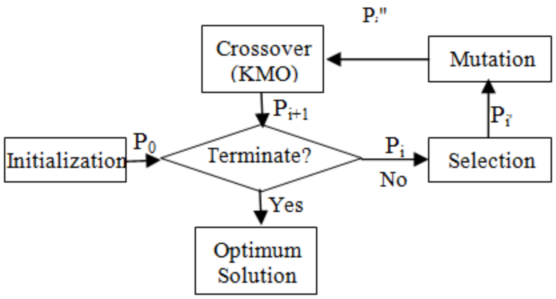

Figure 2. The Flowchart of the GPCMA

5. Experimental Results

For implementation of the system, Matlab [13] software is used. Train data contains 10166 data vectors or 70% of total filtered alert data vectors. The false positive count in the training data set is 4113. Test data includes 30% of the data vectors of labelled alerts; it means 2591 data vectors of true positive, and 1764 data vectors of false positive alerts. The reason of adding the false positive alerts to the test dataset is because IDSs always produce this type of alerts beside the true positive alerts.

In this paper for the training phase the number of clusters (K) assumed 40, the number of chromosomes in population assumed 50, the number of generations is considered 50, crossover probability is 0.8 and finally mutation probability is 0.04 as parameters of GAs.

The parameters of GAs in the test or classify phase are considered 40 for the number of clusters, 40 for the number of chromosomes in population and 40 the number of generations to quit the execution. The probabilities of crossover and mutation are same as training phase.

To evaluate the performance of algorithms three measurement classes are introduced, namely classification, clustering and false positive reduction measurements (Table 1).

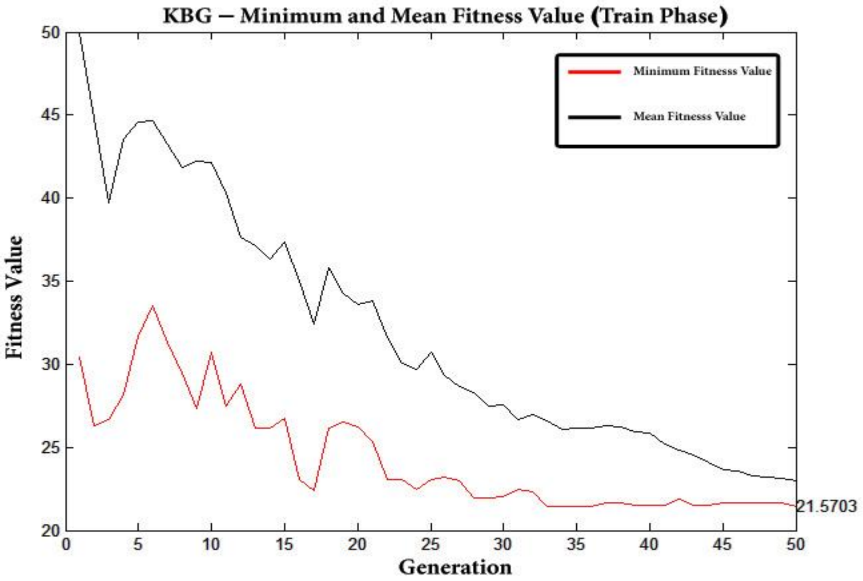

Figure 3. KBG Minimum and Mean Fitness Value in Train Phase

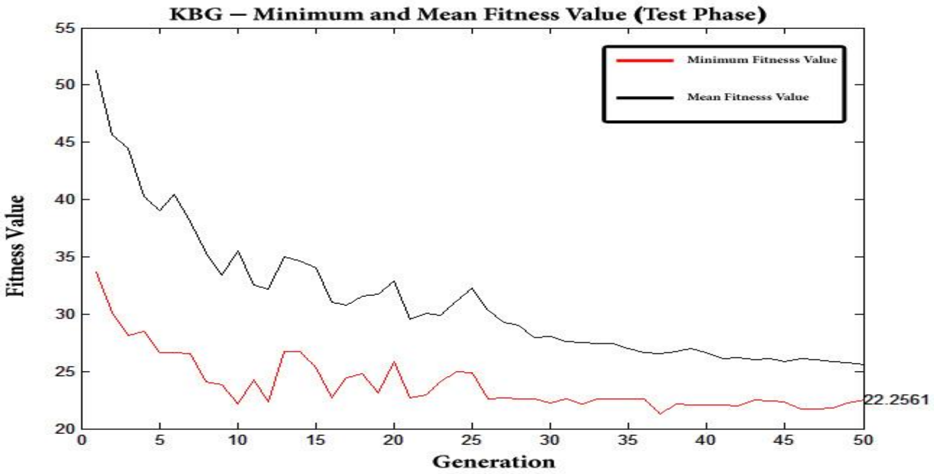

Figure 4. KBG Minimum and Mean Fitness Value in Test Phase

Figure 3 and Figure 4 shows minimum and mean fitness value of KBG in 50 generation.

According to Figures 3 and 4 KBG is convergent, because fitness values in both train and test are descending.

1) Classification Error (ClaE) is the number of alerts that are wrongly classified. 2) Classification Error Rate (ClaER) is the percentage of wrongly classified alerts (20). 3)

Classification Accuracy Rate (ClaAR) is percentage of alerts that are accurately classified as they should be (21). (4) Clustering Error (CluE) is the number of alerts from train data that are wrongly clustered. 5) Clustering Error Rate (CluER) is the percentage of wrongly clustered alerts from train data (22). 6) False Positive Classification Error (FPCE) is the number of false positive alerts that incorrectly classify. 7) False Positive Reduction Rate (FPRR) is percentage of false positive alerts that accurately identified and reduced (23).

$$ClaER=(ClaE \div \text{Total Number of Alerts Observed}) \times 100 \quad (20)$$

$$ClaAR=100-ClaER \quad (21)$$

$$ClaER=(CluE \div \text{Total Number of Alerts Observed From Train Data}) \times 100 \quad (22)$$

$$FPRR=100-(\text{The Number of FP Alerts that Accurately Identified} \div \text{Total Number of FP Alerts Observed}) \times 100 \quad (23)$$

Table 1 shows the average results of eight running of genetic algorithms. As Table 1 shows KBG has best performance over all other seven algorithms.

Table 1. Proposed System Performance Metrics

Algorithm	ClaE	ClaER	ClaAR	CluE	CluER	FPCE	FPRR
KBG	48	2.2	97.8	131	1.1	21	99.2
GA	1218	27.97	72.03	186	1.83	844	52.15
GKA	1011	24.8	75.2	375	3.71	651	62.11
IGA	306	7.03	92.97	564	5.55	84	95.24
FGKA	314	7.21	92.79	772	7.59	44	97.51
GFCMA	148	3.40	96.60	180	1.77	44	97.51
GPCMA	91	2.09	97.91	380	3.74	70	96.03
GFPCMA	148	3.40	96.60	186	1.83	44	97.50

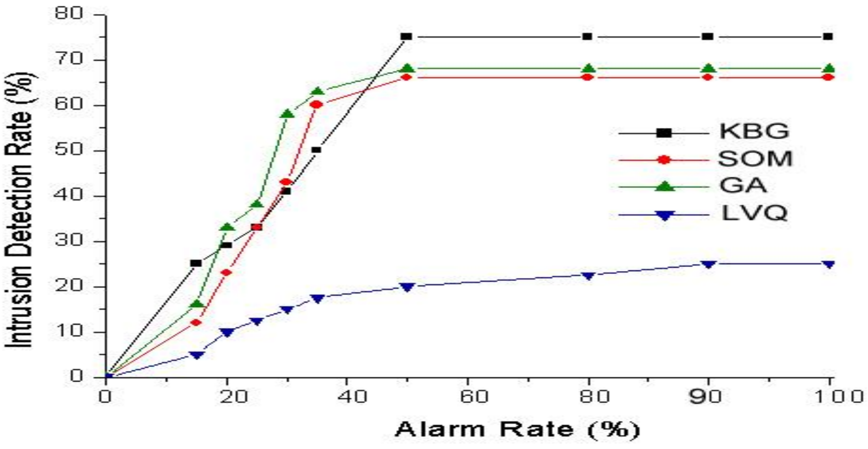

Figure 5. KBG Detection Rate Comparison with Other Algorithms

Figure 5 shows the comparison of four famous clustering techniques: LVQ, SOM, GA and the proposed algorithm KBG.

6. Future Works

A system based on KBG is presented in this paper which can cluster and classify the alerts with high accuracy. This system is also able to reduce the number of false positive alerts considerably.

The trained KBG by alerts with various types of attacks can be a suitable tool to classify the alerts and reduce the false positive alerts in the alert management systems if proper filtering process and pre-processing is executed.

References

[1] H. Debar, M. Dacier and A. Wespi," Towards a taxonomy of intrusion-detection systems", COMPUT. NETWORKS, vol. 31, Iss. 8, **(1999),** pp. 805-822.

[2] H. Debar and A. Wespi, "Aggregation and correlation of intrusion detection alerts", Proc. of the 4th Int. Symp. On Recent Advances in Intrusion Detection, **(2010),** pp. 87–105.

[3] F. Cuppens, "Managing alerts in a multi-intrusion detection environment", Proceedings of the 17th Annual Computer Security Applications Conference on, **(2001),** pp. 22-31.

[4] E. MIRADOR, "Mirador: a cooperative approach of IDS", European Symposium on Research in Computer Security (ESORICS). Toulouse, France, **(2000)**.

[5] J. Wang, H. Wang and G. Zhao, "A GA-based Solution to an NP-hard Problem of Clustering", Security Events. IEEE, pp. 2093- 2097.

[6] J. Wang and B. Cui, "Clustering IDS Alarms with an IGA-based Approach", ICCCAS, **(2009),** pp. 586-591.

[7] A. A. A. Ahrabi, A. H. Navin, H. Bahrbegi, M. K. Mirnia, M. Bahrbegi, E. Safarzadeh and A. Ebrahimi, "A New System for Clustering and Classification of Intrusion Detection System Alerts Using Self-Organizing Maps", International Journal of Computer Science and Security (IJCSS), vol. 4, Iss. 6, **(2010),** pp. 589 – 597.

[8] H. Bahrbegi, A. H. Navin, A. A. A. Ahrabi, M. K. Mirnia and A. Mollanejad, "A new system to evaluate GA-based clustering algorithms in Intrusion Detection alert management system", Nature and Biologically Inspired Computing (NaBIC), Second World Congress on, **(2010),** pp. 115–120.

[9] "MIT Lincoln Lab., DARPA Intrusion Detection Evaluation Datasets", **(1998)** Available: http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/1998data.html".

[10] "Snort: The open source network intrusion detection system", Available: http://www.snort.org.

[11] S. T. Brugger and J. Chow, "An Assessment of the DARPA IDS Evaluation Dataset Using Snort", UC Davis Technical Report CSE-2007-1, Davis, CA, **(2007).**

[12] "Snort Manual", http://www.snort.org/assets/82/snort_manual.pdf.

[13] "Matlab Software", http://www.mathworks.com.

[14] H. Debar and A. Wespi, "Aggregation and Correlation of Intrusion-Detection Alerts", Proceeding RAID '00 Proceedings of the 4th International Symposium on Recent Advances in Intrusion Detection, **(2001),** pp. 87-105.

[15] K. Krishna and M. Murty, "Genetic K-means algorithm", IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics, **(1999),** pp. 433-439.

[16] L. Fuyan, C. Chouyong and L. Shaoyi, "An Improved Genetic Approach", International Conference on Neural Networks and Brain, **(2005),** pp. 641-644.

[17] Y. Lu, S. Lu, F. Fotouhi, Y. Deng and J. S. Brown, "FGKA: a Fast Genetic K-means Clustering Algorithm", Proceeding of the ACM Symposium on Applied computing (SAC), Nicosia, Cyprus, **(2004),** pp. 622-623.

[18] D. E. Goldberg, "Genetic Algorithms in Search, Optimization and Machine Learning", Addison-Wesley, New York, **(1989).**

[19] Z. Michalewicz, "Genetic Algorithms + Data Structures = Evolution Programs", Springer, New York, **(1992)**.

[20] J. L. R. Filho, P. C. Treleaven and C. Alippi, "Genetic algorithm programming environments", IEEE Computer Society Press, vol. 27, issue 6, **(1994)**, pp. 28-43.

[21] M. Saha, "Genetic Algorithm and Simulated Annealing based Approaches to Categorical Data Clustering", Proceedings of the International Multi Conference of Engineers and Computer Scientists, vol. 1, **(2008),** pp. 534-539.

[22] B. Aldenderfer, "Cluster Analysis", Sage Publications, **(1984),** page 7.

[23] R. Krovi, "Genetic algorithm for clustering: A preliminary investigation", Proc. 25th Hawaii Internat. Conf. on Systems Sciences, **(1992)**, pp. 540–544.

[24] U. Maulik and S. Bandyopadhyay, "Genetic algorithm-based clustering technique", Elsevier, The Journal of the Pattern Recognition Society, **(2000),** pp. 1455-1465.

[25] J. A. Hartigan, "Clustering Algorithms", Wiley, 1975, ISBN 0-471-35645-X.

[26] M. Saha, "Genetic Algorithm and Simulated Annealing based Approaches to Categorical Data Clustering", Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 1, **(2008),** pp. 534-539.

[27] K. Timm, “Strategies to reduce false positives and false negatives in NIDS”, Security Focus Article, available online at: http://www.securityfocus.com/infocus/1463, **(2009)**.

[28] M. J. Ranum, “False Positives: A User’s Guide to Making Sense of IDS Alerts”, ICSA Labs IDSC, **(2003)**.

[29] S. Zanero, “Flaws and Frauds in the Evaluation of IDS.IPS Technologies”, first accessed on 21.09.07, **(2007),** http:// www.first.org/conference /2007/papers/zanero-stefano-paper.pdf.

[30] J. Snyder, Taking Aim: “Target–Based IDS Squelch Network Noise to pinpoint the alert you really care about”, Information security Magazine, **(2004)** January.

[31] V. Mattord, “Principles of Information Security”, Course Technology. **(2008),** pp. 290–301, ISBN 978-1-4239-0177-8.

[32] S. X. Wu and W. Banzhaf, “The use of computational intelligence in intrusion detection systems: A Review”, Applied Soft Computing Journal, vol. 10, **(2010)**.

[33] S. Wu and E. Yen, “Data mining-based intrusion detectors”, Expert Systems with Applications, vol. 36, **(2009)**.