Enhanced Solutions for Misuse Network Intrusion Detection System using SGA and SSGA

Sabah A. Jebur[†] and Hebah H. O. Nasereddin [††]

Middle East University, Amman, Jordan

Summary
One of the most widely acknowledged purposes of using the internet is data transfer; it is an essential way of communicating personal and sensitive data. Therefore, the need for protecting such data against hackers and intruders is at most importance. Many security systems were built for this purpose; Intrusion Detection Systems (IDS) are one of those systems. The main function of IDS is to monitor the incoming connections and detect attacks. In this paper, the researcher presented two models of IDS. In the first model, the Simple Genetic Algorithm (SGA) was used to build IDS, while in second model; Steady State Genetic Algorithm (SSGA) was used to build IDS. The evaluations and the experiments were performed using the NSL-KDD dataset. The experimental results demonstrated that performing an IDS using SGA gives higher performance results than using SSGA according to the value of Detection rate (DR) and number of new generated rules, also the training time for SGA experiments is shorter than the training time for SSGA. On other hand, SSGA based IDS achieved an average of False Positive Rate (FPR) that was relatively better than SGA based IDS.

Key words:
Intrusion Detection System, Simple Genetic Algorithm, Steady State Genetic Algorithm

1. Introduction

Attacks on the computer resources are becoming an increasingly serious Problem nowadays. Despite different techniques have been developed and deployed to protect computer systems against network attacks, securing data communication over internet and any other network are still under threat of intrusions [1]. Intrusion is the set of actions that attempts to compromise integrity, confidentiality or availability of network resources [2]. Firewalls, access controls, and authentication facilities play a major role in countering intrusions. IDS is often used as another line of defense to monitor and analyze system events in order to detect intrusions by assuming that the intruder's behavior differs from the authorized user's behavior. IDS is a security system that monitors and analyzes system events for the purpose of finding, and providing real-time or near real-time warning of, attempts to access system resources in an unauthorized manner [3]. The motivations of using IDS for intrusions is detected quickly and ejected from the system before causing any damage, it is an action to prevent intrusions. There are generally two accepted categories of intrusion detection techniques: Misuse detection and Anomaly detection. Misuse detection technique looks for sequences of known events to identify attacks. It depends on a definite set of rules or attack patterns that can be used to detect the intrusion [4]. Anomaly detection technique depends on collecting information about the behavior of authorized users over a period of time by analyzing incoming audit records to identify deviation from an average behavior [4]. An IDS also is divided into two groups depending on where they are looking for an intrusive behavior: Network-based IDS (NIDS) and Host-based IDS (HIDS). HIDS adds a specialized layer of security software to vulnerable or sensitive systems such as database servers and administrative systems. HIDS monitors the activity on the system in a variety of ways to detect suspicious behavior. It can detect both external and internal intrusions [3]. While NIDS monitors network traffic for particular network segments or devices. An NIDS analyzes the traffic packets in real time or near to real time to identify suspicious activity [5]. The most of existent IDSs face a number of challenges such as low DR and high FPR and therefore prevent authorized users from accessing the network resources, these problems occur because of the sophistication of the attacks and their intended similarities to normal behavior [6]. To overcome these problems, IDS must be implemented by using smart methods based on artificial intelligence techniques to detect the attacks. One of the important approaches of artificial intelligence used to detect Intrusion is Genetic Algorithm (GA).

2. Genetic algorithm (GA)

GA is the powerful stochastic algorithm which is applied in machine learning and optimization problems to solve complex problems. It is based on the principles of natural selection and natural genetics inspired by Darwin‘s principles in optimizing the chromosome population of candidate solutions. GA maintains a population of individuals and probabilistically modifies the population

Manuscript received May 5, 2015
Manuscript revised May 20, 2015

using genetic processes, with the intent of seeking a near optimal solution to the problem [7]. It starts with a population of individuals randomly sampled over the search space. Using fitness function, each individual is associated with a fitness value that reflects its quality. GA tries to improve the quality of the individuals by making the population evolve. This evolution is achieved using information exchanges between individuals in order to create new ones or modify the existing ones using genetic operators such selection, crossover and mutation [8]. There are two categories of GA will be discussion in this paper: Simple Genetic Algorithm (SGA) and Steady State Genetic Algorithm (SSGA). SGA creates new chromosomes (offspring's) from current chromosomes (parents) using the genetic operators (Crossover and Mutation). These new chromosomes replaced previous chromosomes to form new population for the next generation, where all of the population undergoes transformation at each generation [9]. SSGA automatically includes the current best chromosomes in the next generation, and only the poorest chromosomes will be replaced. Therefore, SSGA allows some chromosomes to survive over time due to the Replacement process, because it allows some part of the current population to be carried to next generation based on their fitness value [9]. SGA and SSGA differ significantly in how individuals survive over time, how chromosomes are replaced, and how often they may reproduce. The replacement strategy likely to have a significant effect in producing advanced chromosomes due to the fact that it differs in SGA from SSGA. This paper will verify the power of SGA versus SSGA in intrusion detection field, by calculating the training time, number of new generated rules, Detection Rate (DR), and False Positive Rate (FPR).

3. GA Components

The main components of a GA are:

3.1 Population

It is array of chromosomes will be randomly generated at the start of GA to cover the range of possible solutions (the search space), where each chromosome represents potential solution of the problem to be solved. The nature of the problem determines the population size [15].

3.2 Evaluation

The evaluation process is a very important measure to calculate the goodness of a chromosome. Fitness function is the heart of all Genetic Processes. It evaluates the performance of all chromosomes in the population, where a chromosome with high fitness value has a high probability to be selected in the selection stage [10].

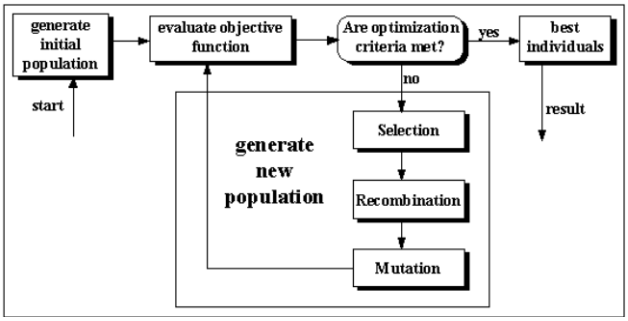

Fig. 1 Genetic Algorithm Structure.

3.3 Encoding

Encoding is one of the main processes in GA to represent the data into some of the encoding formats. Various encoding methods have been created for particular problems to provide effective implementation of genetic algorithms. The encoding methods can be classified into Binary Encoding, Integer or Literal Permutation Encoding, and Real Number Encoding [10].

3.4 Selection

In this process, multiple chromosomes are selected from the current population based on their fitness value to produce successive generations. The better chromosomes have more chance of being selected and can be selected more than once to reproduce into the next generation. There are several schemes for the selection process, such as Roulette Wheel Selection (RWS), Stochastic Universal Sampling (SUS), Rank Selection, Elitism Selection, and Tournament Selection [11].

3.5 Crossover Operator

The crossover operator decides which genes of the parents should be swapped to generate the offspring. There are several crossover operators, such as One-point crossover**,** Two-point crossover, N-point crossover, and Uniform crossover [12].

3.6 Mutation Operator

It is used to produce new chromosomes or modify some features of them depending on some small probability value. The objective of this operator is to prevent falling of all solutions in population into a local optimum of solved problem. There are several types of mutation methods, such as Flip bit, Boundary, Inversion, Insertion, Displacement and Non-uniform mutation [13].

3.7 Replacement

It is a process performed on the worst individuals to be replaced by better new individuals; this process is used only with SSGA. There are two methods of Replacement, Binary Tournament Replacement (BTR) and Triple Tournament Replacement (TTR) [14].

3.8 Stopping Criteria

This process defines the conditions that the search process terminates. Typical stopping criteria include: maximum number of generations reached, if Successive iterations no longer produce better results, If there are no additional new solutions will be produced, and terminate if the optimal solution has been discovered [15].

4. Literature Review

Hoque et al. [16] presented and implemented a NIDS using SGA to efficiently detect various types of network intrusions. This approach used evolution theory to information evolution in order to filter the traffic data and thus reduce the complexity and also used the standard KDD99 benchmark dataset to implement and measure the performance of their system. The authors used only the numerical features, both continuous and discrete, also used the standard deviation equation with distance to measure the fitness of a chromosome. They got the following Detection Rate results (Probe: 71.1%), (DoS: 99.4%), (U2R: 18.9%) and (R2L: 5.4%). Torkaman et al. [17], on the other hand, designed a hybrid approach for modeling HIDS combines anomaly and misuse detection, based on two-layer fuzzy Genetic algorithm and neural network which uses simple data mining techniques to process the web application traffics, Two-layer fuzzy Genetic algorithm and neural network are applied respectively as anomaly and misuse detection. One of the advantages of this algorithm is that, it can support multiple attack classification according to Open Web Application Security Project (OWASP). This search used The HTTP dataset CSIC 2010 which is generated automatically and contains 36,000 normal requests and more than 25,000 anomalous requests. The proposed model is able to detect critical vulnerabilities based on OWASP standard. In [18] researchers developed a real-time detection approach for detecting anomaly attacks. They used packet sniffer to sniff network packets in every 2 seconds and preprocessed it into 12 features and used Fuzzy Genetic algorithm to classify the network data. The fuzzy rule is a supervised learning technique and genetic algorithm make fuzzy rule able to learn new attacks by itself. The output can be categorized into DoS and Probe. The network dataset used for training and testing is collected in the actual network environment in their research laboratory. The result shows that this algorithm has over 97% of DR, less than 1% of FPR and less than 3 seconds (for data preprocessing and detection) to issue the alert message after an attack has arrived. Naoum et al. [14] enhanced SSGA based IDS for detecting misuse attacks by comparing Replacement methods. The research demonstrates that the Triple Tournament Replacement produces more accurate results than Binary Tournament Replacement according to the value of DR and the number of new chromosomes. It got the average of DR which is equal to 88.25%, and the average of FPR that is equal to 1.48%. The experiments and evaluations are performed by using 10% of the whole KDD99 dataset.

5. Proposed Models

This study aims at verifying the power of SGA against SSGA in Intrusion Detection field. This is achieved by building two models of IDS: SGA based IDS and SSGA based IDS, then measuring the performance of each model by calculating the training time, number of new generated rules, Detection Rate (DR), and False Positive Rate (FPR). Figure (2) and figure (3) display our proposed models.

5.1 Processing Phase

This phase starts with importing the dataset and the encoding of features, then selecting the dataset with the reduced features for each attack, then filtering the duplicated rules by eliminating redundant ones, this phase include:

NSL-KDD Dataset: is a benchmark used to evaluate the system efficiency by measuring its performance; it is publicly available on (http://nsl.cs.unb.ca/NSL-KDD/). It

consists of training dataset and testing dataset. The training dataset contains 125,973 records, while the testing dataset contains 22,544 records.

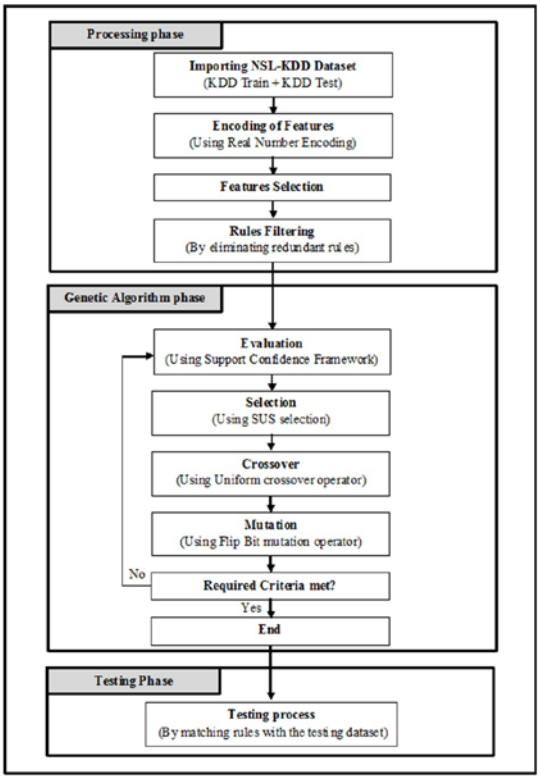

Fig. 2 The structure of SGA based IDS.

Encoding of Features: In NSL-KD dataset each rule contains 41 features, three of which are of string values, these features are Protocol Type, Flag and Service. Real Number Encoding is used to encode those features as following:

- **Protocol type feature** has three different values; TCP, ICMP and UDP. So, TCP is encoded by number 1, ICMP is encoded by number 2 and UDP is encoded by number 3.
- **Flag feature** has 11 different values; each value was represented by a positive number, such as "REJ" value which was represented by number 1, "OTH" by number 2 … etc.
- **Service feature** has 70 different values. The researcher used positive numbers to represent these values, such as "BGP" value which was represented by number 1, "SQL_NET" by number 2 … etc.

Features Selection: Each record in NSL-KDD dataset is described by 41 features, using all these features to generate rules is a time consuming process. So, the most significant features should be selected to represent each attack category. There are several studies that have proposed different features sets to represent different type of attacks. In order to select the most significant features to be used in this research, the researcher performed the testing process over the testing dataset using the selected features sets by [19], [20], and [21]. According to testing results, The researcher selected the features sets shown in table (1) to use them in this research.

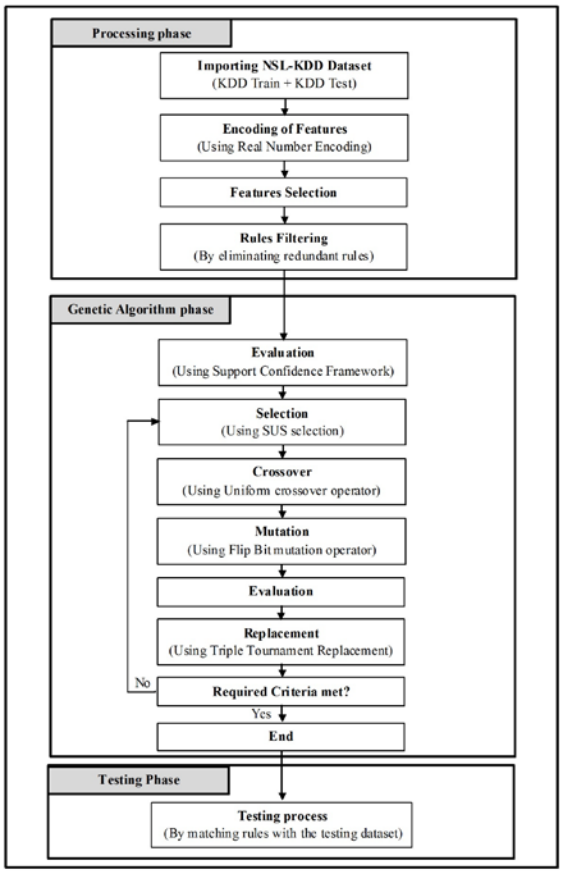

Fig. 3 The structure of SSGA based IDS.

Table 1: The selected features sets for this esearch

Attack name	Features sets
Dos	F5, F38, F3
Probe	F3, F12, F27, F31, F35
U2R	F1, F2, F3, F10, F16
R2l	F3, F4, F6, F9,F11, F22, F25, F33, F37, F38

Rules Filtering: After selecting the training dataset with the reduced features for each attack, many duplicated rules have appeared. These duplicated rules are unnecessary and keeping them could slow the work down. So, training records were filtered by eliminating the redundant rules. After analyzing the training dataset, each rule will be

represented in: *if* condition *then* action. The condition part refers to the features of the network connection, the result might be TRUE, if the incoming connection matched the rules in dataset, or it could be FALSE if there was some mismatching. The action part refers to the attack name and will be specified only if condition is true.

5.2 Genetic Algorithm Phase

This phase aims at generating new rules to be used for detecting attacks in testing dataset. In this phase, SGA based IDS applied Evaluation, Selection, Crossover and Mutation processes, while SSGA Based IDS performed Evaluation, Selection, Crossover, Mutation as well as Replacement processes.

Evaluation: Support Confidence Framework were adopted as the fitness function in this research in order to evaluate each individual in the training dataset. This fitness function is developed by [22]. If a rule is represented as (if A then B), then:

Support = |A and B| / N

Confidence = |A and B| / |A|

Fitness = w1 × Support + w2 × Confidence

Where:

N = number of connections in training dataset.

|A| = number of records matching the condition A.

|A and B| = number of records matching the rule (if A then B).

w1, w2 = weights to balance the two terms.

Selection: This study used Stochastic Universal Sampling (SUS) as selection method. It is implemented by obtaining N equally spaces pointers by generating single random number between [0, FV] as pointer1, and then adding (AF) to generate next pointers, and so on. Where N is the number of required selections, and AF is the average of fitness value in the population. The individual who has a fitness value that spans the positions of the pointers is selected.

Crossover: Uniform crossover operator is used as a crossover method in this research. It is considered the most powerful crossover because all genes have equal probability to be swapped in order to gain a high diversity in population. It is implemented by randomly exchange genes between two parents where the offspring will have 50% of the first parent's genes and another 50% from the second parent's genes [23].

Mutation: Flip bit mutation operator is used as a mutation method; it is performed by randomly selecting gene and makes its value equal to a random number of specific range. The probability of mutation rate used in the experiments is 10%, one individual among every 10 will undergo a mutation process.

Replacement: This process is used only in SSGA based IDS model. Triple Tournament Replacement (TTR), used in this research, is implemented by comparing three generations among each other, while the rules with highest fitness values will be selected and stored in the rules pool.

Stopping Criteria: the GA phase will be stopped If there were no additional new rules to be produced.

5.3 Testing Phase

After applying Genetic Algorithm process over the training dataset and generating new rules. These generated rules will be used to detect attacks from the testing dataset for evaluating the proposed models. The evaluation is implemented by calculating DR and FPR for each proposed model, where:

DR = No. of Detected Attacks / No. of Total Attacks.

FPR = No. of False Alarms / No. of Normal Alarms.

6. Experimental Results

In this study, the researcher conducted two types of experiments; on the first type, SGA was used for obtaining rules in 20 sub attack types. Then, he tested these rules with testing dataset. On the second type, the researcher applied the same method as in the first experiment but used SSGA instead of SGA. Table (2) and table (3) illustrate experimental results from SGA based IDS and SSGA based IDS, respectively.

Table 2: Experimental results from SGA based IDS

Attack name	No. of new generated rules	Training time	DR	FPR
Dos	4037	02:20:40	91.78%	1.92%
Probe	127927	07:51:20	93.58%	12.14%
U2R	7662	00:32:33	81.08%	3.92%
R2l	100083	28:34:20	87.54%	2.86%

Table 3: Experimental results from SSGA based IDS

Attack name	No. of new generated rules	Training time	DR	FPR
Dos	1342	23:23:20	91.36%	1.92%
Probe	59268	28:42:04	81.83%	10.19%
U2R	524	00:46:00	40.54%	3.53%
R2l	39630	33:00:52	76.40%	3.00%

7. Conclusion

This study proposed two models of Intrusion Detection Systems to detect network intrusions. In the first model, an Intrusion Detection System is built using Simple Genetic Algorithm (SGA based IDS). While the second model, an Intrusion Detection System is built using Steady State Genetic Algorithm (SSGA based IDS). These models were implemented and evaluated using NSL-KDD intrusion detection dataset. In order to determine which feature set is the most suitable one for each attack category, the author compared published studies regarding this topic and modeled a hybrid feature set containing the best available feature sets. The training process showed that the training time for SSGA was much more than the time required for SGA, despite that, the numbers of new generated rules using SGA are more than those using SSGA. The experimental results demonstrated that SGA based IDS achieved more accurate results than SSGA based IDS according to DR where it produced an average of DR equal to 88.5%, while SSGA based IDS produced a result of DR equal to 72.53%. The results of FPR using SGA based IDS, achieved higher results than SSGA based IDS in R2L attacks, but it got lower results than SSGA based IDS in Probe and U2R attacks while achieving equal result in Dos attacks.

References

[1] B. Abdullah, I. Abd-Alghafar, G. Salama and A. Abd-Alhafez, “Performance evaluation of a genetic algorithm based approach to network intrusion detection system”, Proceedings of the International Conference on Aerospace Sciences and Aviation Technology. 2009.

[2] A. Ojugo, A. Eboka, O. Okonta, R. Yoro and F. Aghware, "Genetic algorithm rule-based intrusion detection system (GAIDS)", Journal of Emerging Trends in Computing and Information Sciences, Vol.3, No.8, ISSN: 2079-8407, pp. 1182-1194, 2012.

[3] W. Stallings and L. Brown, "Computer Security: Principles and Practice", Upper Saddle River, NJ: Prentice Hall. 2008.

[4] F. Shiri, B. Shanmugam and N. Idris, " A parallel technique for improving the performance of signature-based network intrusion detection system", In Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on (pp. 692-696), 2011

[5] K. Prasad and S. Borah, "Use of Genetic Algorithms in Intrusion Detection Systems: An Analysis", International Journal of Applied Research and Studies (iJARS), Vol.2, Issue 8, ISSN: 2278-9480, 2013.

[6] E. Shaveta, A. Bhandari and K. Saluja, "Applying Genetic Algorithm in Intrusion Detection System: A Comprehensive Review", Association of Computer Electronics and Electrical Engineers, 2014.

[7] P. Guo, X. Wang and Y. Han, “The enhanced genetic algorithms for the optimization design”, 3rd International Conference on Biomedical Engineering and Informatics (BMEI), Vol. 7, pp. 2990-2994. IEEE, 2010.

[8] A. O. Adewumi, "Some improved genetic-algorithms based heuristics for global optimization with innovative applications", Master thesis, University of the Witwatersrand, Johannesburg, South Africa, 2010.

[9] M. Mehra, M. Jayalal, A. Arul, S. Rajeswari, K. Kuriakose and S. Murty, “Design and Development of Genetic Algorithm for Test Interval Optimization of Safety Critical System for a Nuclear Power Plant”, In Online Proceedings on Trends in Innovative Computing, Intelligent Systems Design and Applications Conference, Kochi, India, 2012.

[10] B. Dhak and S. Lade, "An evolutionary approach to intrusion detection system using genetic algorithm", International Journal of Emerging Technology and Advanced Engineering, Vol. 2, Issue 12, ISSN: 2250-2459, 2012

[11] T. Pencheva, K. Atanassov and A. Shannon, "Modeling of a stochastic universal sampling selection operator in genetic algorithms using generalized nets", In Proceedings of the Tenth International Workshop on Generalized Nets, Sofia (pp. 1-7), 2009

[12] G. Soon, T. Guan, C. On, R. Alfred and P. Anthony, “A comparison on the performance of crossover techniques in video game”. IEEE International Conference on Control System, Computing and Engineering, 29 Nov. -1 Dec. 2013, Penang, Malaysia. IEEE, 2013.

[13] B. Hasan and M. Mustafa, “Comparative Study of Mutation Operators on the Behavior of Genetic Algorithms Applied to Non-deterministic Polynomial (NP) Problems”. Second International Conference on Intelligent Systems, Modeling and Simulation, (pp. 7-12). IEEE, 2011.

[14] R. Naoum, S. Aziz and F. Alabsi, "An Enhancement of the Replacement Steady State Genetic Algorithm for Intrusion Detection", International Journal of Advanced Computer Research, Vol.4, No.2, Issue 15, ISSN: 2249-7277, 2014.

[15] M. Kumar, M. Husian, N. Upreti and D. Gupta, "Genetic algorithm: review and application", International Journal of Information Technology and Knowledge Management, Vol.2, No.2. PP. 451-454, 2010.

[16] M. Hoque, A. Mukit and A. Bikas, "An Implementation of Intrusion Detection System Using Genetic Algorithm", International Journal of Network Security and its applications, Vol.4, NO.2, 109-120, 2012.

[17] A. Torkaman, G. Javadzadeh and M. Bahrololum, “A hybrid intelligent HIDS model using two-layer genetic algorithm and neural network”. 5th Conference on Information and Knowledge Technology (IKT), (pp. 92-96). IEEE, 2013.

[18] P. Jongsuebsuk, N. Wattanapongsakorn and C. Charnsripinyo, “Network intrusion detection with Fuzzy Genetic Algorithm for unknown attacks”. International Conference on Information Networking (ICOIN), (pp. 1-5). IEEE, 2013.

[19] S. Mukkamala, A. Sung and A. Abrham, "Modeling Intrusion Detection System using Linear Genetic Programming Approach", Proceeding IEA/AIE 17th International Conference on Innovations in Applied

Artificial Intelligence, PP 633- 642, ISBN: 3-540-22007-0, 2004.
[20] T. Chou, K. Yen and J. Luo, "Network Intrusion Detection Design Using Feature Selection of Soft Computing Paradigms". International journal of computational intelligence, Vol.2, No.11, pp. 196-208, 2008.
[21] F. Amiri, M. Yousefi, C. Lucas, A. Shakery and N. Yazdani, "Mutual information-based feature selection for intrusion detection systems" Journal of Network and Computer Applications, Vol. 34, Issue 4, pp 1184-1199, 2011.
[22] M. Wong and K. Leung, "Data mining using grammar based genetic programming and applications", Kluwer Academic Publishers, Netherlands, 2000.
[23] X. Hu and E. Di Paolo, "An efficient genetic algorithm with uniform crossover for air traffic control", Computers & Operations Research, 36(1), 245-259, 2009.

Sabah A. Jebur received the Bachelor degree in Computer Science from Mustansiriyah University, Iraq in the year 2003. Presently he is a Master student of Computer Information Systems in Middle East University, Jordan. His research interests include Genetic Algorithms, Evolutionary Algorithms, and Network Security.

Hebah H. O. Nasereddin, Associate Prof. in Faculty of Information Technology, Middle East University (MEU), Amman, Jordan. Nasereddin is a reviewer for several National and International journals and a keynote speaker for many conferences, general chair for ICNVICT. She is supervising many MSc, and Diploma thesis. Her research is focused on Data warehouse, Data Mining, Cryptography, Steganography, and Big Data. Dr. Nasereddin published in Computer Philosophy and other Computer topics publications. She is Chief Editor and Editor for several Magazines in addition to her participation in project research evaluations.