Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 10, No. 2, May 2019), Pages: 1-16
www.jacr.iausari.ac.ir

# A Parallel Genetic Algorithm Based Method for Feature Subset Selection in Intrusion Detection Systems

**Iran Shokripoor Bahman Bigloo**✉
*Department of Computer Engineering, Kerman Branch, Islamic Azad University, Kerman, Iran*
iran.shokripoor@yahoo.com

**Manuscript ID: JACR-1810-1645** **Received: 2018/10/18; Accepted: 2018/12/17**

**Abstract**

*Intrusion detection systems are designed to provide security in computer networks, so that if the attacker crosses other security devices, they can detect and prevent the attack process. One of the most essential challenges in designing these systems is the so called curse of dimensionality. Therefore, in order to obtain satisfactory performance in these systems we have to take advantage of appropriate pre-processing steps specially the feature subset selection methods. Since the problem of searching for the optimal feature subset has an intolerable complexity, in this paper we propose a genetic-algorithm-based search method for finding the most relevant subset of features. In order to find the most relevant features, the parallel structure of the genetic algorithm along with the distribution factor of the features is used. The fitness value of each feature subset is computed according to performance of the classifier trained with respect to that subset. In order to evaluate the performance of the proposed method, we use the NSL-KDD dataset which benefits from more real-world intriguing records than other intrusion detection data. The results of our evaluation experiments shows that the proposed method outperforms the prior methods.*

***Keywords:*** *Intrusion Detection Systems, Data-Mining, Feature Selection, Genetic Algorithm, Dataset NSL-KDD*

## 1. Introduction

Importance of network security has grown increasingly and the Intrusion Detection Systems (IDS) has been developed rapidly so that they help reach this goal and more security. Intrusion detection and prevention include three main duties of data collection, data analysis, and response operations. In data collection, the system collects its required information such as access to various under-supervision files or information on network performance. In host-based systems, data are collected based on the internal sources of the host which are mostly in the operating system level. On the other hand, in network-based systems, by analyzing the packets passing through the network, the required parameters can be delivered to the analysis section to detect the intrusion. It is a serious issue in IDS performance to determine exactly which data should be collected. Most operating systems collect the security related events which can be a good source of information for IDS. When the idea of using such systems was discussed, due to the high processing load, it was only welcomed by the military and business environments.

As the design and production of the hardware have been evolved significantly, it is possible to use this technology and idea for a wide range of computer systems. There are numerous intrusion detection systems in which the main challenge is to enhance the efficiency [1]. Most current IDSs use all existing parameters in network packets for evaluating and exploring the attack patterns, while some of these parameters are redundant and irrelevant. Using all parameters makes the detection process lengthy decreasing the IDS efficiency. In fact, the main challenge in IDSs is the high volume of data [2]. Additionally, in view of high traffic, the decreased false alarm rate in the intrusion detection system is of high importance.

All intrusion detection systems are able to generate an alarm in case of any intrusion in the network. But, due to the high volume of alarms generated by such systems and also false alarms generation, these systems are not able to manage and analyze the generated alarms. Todays, most approaches to intrusion detection focus on the issue of selection or extraction of important features. A number of features with the best accuracy in the intrusion detection is obtained experimentally in various datasets. But, selecting the features may lead to loss of a part of data. Therefore, with regards to the different data related to the network, estimating the optimum number of features is considered one of the important challenges in such a system.

In this study, attempt is made to select those features as the basic features that increase the accuracy of intrusion detection and also, we try to incorporate the effect of features which have not been selected in the intrusion detection system. For this purpose, by using the parallel genetic algorithms, we are going to select those features which play the most important role in classifying data, and to obtain its optimum number automatically. In present study, the data-mining technique is used for classifying the selected features. By using the data-mining method, one can apply the complete management on the alarms generated by the intrusion detection systems. Also, the number of alarms generated by such systems can be reduced and make their analysis easier through ignoring the false alarms.

The intrusion detection methods include the abnormal behavior detection and abuse detection (based on signature) [3]. There are various types of architecture for intrusion detection systems which can be generally classified into three categories: host-, distributed- and network- based intrusion detection systems. The intrusion detection systems are responsible for identifying and detecting any unauthorized use of system, abuse or damage imposed by both internal and external users. IDSs are established in hardware and software forms with the own advantages and disadvantages. Both of the speed and accuracy are the advantages of hardware systems, and their lack of security failure by the attackers is the other capability of such systems. Generally, three main functions of IDS are supervision and evaluation, exploration and reaction.

This study is organized as below. In Section 2, some of the most recent works are discussed. In the section chapter, the proposed genetic algorithm based intrusion detection system (IDS) is presented and also, the necessary operators are given for selecting the most effective features of the system. Results obtained from the system evaluation are presented in Section 4. Finally, the conclusions and suggestions are discussed in Section 5. In this work, first, we will review the reduction in penetration detection. In Section 3, the proposed intrusion detection system based on the genetic algorithm is described and the necessary operators to select the most effective features

are presented. The evaluation results of the proposed system are presented in Section 4. Finally, the conclusions and suggestions are presented in Section 5.

## 2. Review of Literature Research History

Intrusion detection and prevention is one of the main mechanisms for meeting the security of computer systems and networks. The IDSs are taken into account as separate systems, but such systems are applied as subsystems of network equipment, operating systems, and even services. Many researches have been conducted in the field of reducing the features of intrusion detection systems which among them, some methods are based on artificial intelligence or data-mining.

In 2008, a genetic algorithm based method was presented for decocting the intrusion on KDDCUP99 database [4]. This method uses a machine learning approach with the genetic algorithm for extracting the features. A special type of intrusion is identified by producing a series of rules based on each rule. In another study, K-means clustering hybrid learning approach and simple Bayesian classification [5] have been given. Clustering attributes all data to the related groups before using the classification. In this study, KDDCUP99 data are used for training which shows the better performance results in terms of the accuracy and speed in intrusion detection with proper false alarm rate.

Many studies have been done to reduce the dimensions of the features in penetration detection systems which among them, some methods are based on the artificial intelligence and the some others are based on the data mining. Most of the data mining works are categorized into the choosing of the features and models. In 2008, a new method based on the genetic algorithm was presented to detect the intrusion on the KDDCUP99 database [4]. This method uses a machine learning approach with a genetic algorithm to extract the features. The result of this research is a set of rules which based on them, a specific type of penetration is identified. Due to the high number of the extracted rules, the complexity of this method makes it inefficient in the online systems. In another study, a clustering learning approach and a simple bundle categorization were proposed in which clustered before the data classification model was built [5].

The use of clustering features identifies the related features. As a result, the simple binary categorization with the different specification (the select of the features from different clusters) shows the better performance in terms of the accuracy and speed of intrusion detection with the appropriate false alert rate.

Developed an intrusion detection system based on the machine learning [6]. They applied the genetic algorithm with SVM in order to determine automatically the proper set of features. This issue supports the application of the genetic algorithm for feature selection and SVM based classification. In the present study, there is a pre-defined dictionary with the type of attack. Such method presents the most distinct features for any type of attack. The results of this method show that the average value of the selected features in some cases is close to the total number of main features. Developed an intrusion detection system using the genetic algorithm in which an idea is offered for improving the sections of primary population and selection operator in genetic algorithm [7].

In this method, the convergence has been accelerated due to the onset of a genetic algorithm with a high quality primary population. A novel method for selecting the effective features to form the intrusion detection model was given in the research presented [8]. The proposed method for feature selection is calculated in view of the mean features of each class to the whole class. The standard feature selection methods; i.e. Correlation-based Feature Selection (CFS), Information Gain (IG) and Gain Ratio (GR) were used to evaluate the performance of the proposed method. For the classification, a decision tree based model has been used which reports the promising results especially by choosing the weight-based feature. This method was classified using the decision tree based algorithm. In another research, ant colony algorithm has been used for extracting the feature in intrusion detection system [9]. Due to use of the simple subsets for classification, this method has fast implementation capability and low computational complexity. The decreased number of features using the graph and explorative information for updating the pheromone leads to an increased accuracy in intrusion detection and reduced false alarm. Although this technique uses a high threshold level to map out the attributes to the graph and to reduce the complexity of the computations, some of data will be lost.

In [10] investigated the SVM and SOM neural network in intrusion detection system. Analysis of the aforementioned method on two datasets of KDDCUP and DARPA showed that SVM is superior to SOM in terms of efficiency and computational speed. One of the reasons for the weakness of the neural network in the intrusion detection system is the problem with detecting the appropriate size of the network and its weights. The experimental results show the accuracy above 99% for SVM. In another study reported by [11], a hybrid classification model based on the tree algorithms was created to detect the disorder in the network. The detection algorithm performance was evaluated using NSL-KDD database which is the updated version of KDDCUP99. This algorithm aims at clarifying this issue whether 41-feature based input network traffics is inspired by the natural network or is considered an attack. A hybrid NBTree algorithm consisting of decision tree classification and Naïve-Bayes classification is used in the present study. Detection accuracy of 89.24% is obtained using a combination of random tree and NBTree algorithms on the basis of the rule scheme which is better than the single random tree algorithm.

Here, a hybrid NBtree algorithm consisting of the decision tree and the Naive-Bayes grouping has been used. Although this method uses all the features and only provides a hybrid classification model, it shows a good performance with the detection accuracy of 89.24% for the combination of the random tree and NBTree algorithms according to the sum of the rule scheme.

The orientation is made toward to the classification by use of the Artificial Immune System (AIS) with the Population-based incremental learning (PBIL) and collaborative filtering (CF) to detect the intrusion in network. The Artificial Immune System (AIS) is a powerful tool in terms of destroying the antigens and is inspired from the natural immune systems. PBIL exploits the previous experiences in making and evolving the new species by learning and CF idea accommodation for classification. The innovation of this new method lies in combination of three mentioned methods for creating a new classification which is enjoying the incremental learning capability so that it is used in

intrusion detection. Additionally, a mean related antibody hierarchical accommodation is proposed for promoting the AIS performance.

Internet expansion everywhere has enhanced the use of computer systems and networks.

In addition, the computational environments are generally proceeding from the conventional focused computer systems towards the distributed systems. As the distributed systems grow, they are more exposed to new attacks. In a computer system which is based on the network, system security supply in is highly important. With the lack of security, it is possible that the system suffers from the disorder and inefficiency. To cope with the attackers to systems and computer networks by both domestic users and foreign attackers, we investigate a novel approach based on the genetic algorithm in detecting and preventing the intrusion and future of this kind of technology.

In this work, due to the limitation of the most methods in determining the number of effective features for the proposed classification model, an intrusion detection system based on the selection of the effective features has been proposed. The select of the effective features is done to reduce the number of features used in intrusion detection and finally, to reduce computational complexity.

In this paper authors proposed the Causality-based Medical Diagnosis and Treatment System with two main capabilities: Diagnosis and Treatment. It can predict the disease of the new patients in tandem with suggesting treatments for currently known patients. Authors have implemented our Medical_CREAM method for finding causal actions and tested on different causal networks. The results have shown that our method is more successful in finding cost-effective actions than current state of the art method –Yang method- in action mining domain [12]. This paper presented some significant changes to improve efficiency of the basic Genetic algorithms for scheduling tasks in heterogeneous systems based on TRIZ. Our Algorithm, TRIZGA-TS, generates an initial population embedded with non-random Schedules that is in contrast to basic genetic algorithms. The results reveal superior performance of TRIZGA-TS over HEFT-T, HEFT-B and CPOP. Also, it offers better scheduling with lowered repetitions in Comparison with BAG, SA and SLPSO as Meta heuristic Approaches [13]. In the present article, a new method for finding the number of optimum cluster heads within an Ad-hoc sensor network is presented using to the Fast and Elitist Multi Objective Genetic Algorithm (NSGA-II).The results indicate the effectiveness of the proposed method in comparison with other common methods in this field [14].

## 3. Proposed Intrusion Detection System

The first step in creating each model based on data-mining techniques is the data pre-processing stage. Pre-processing is carried out for preparing the data to be processed and also, to improve the quality of the real data. This stage includes data normalization and mixture. In the next stage, the desired features are extracted using the parallel genetic algorithm. In some datasets, the number of features is high and it is possible that some of these features do not play a role in data classification. Therefore, it is necessary to select a subset consisting of the best features. The method considered for features selection acts on the basis of features distribution measurement. The presented method can work on datasets with different dimensions. After selecting the features, two

classifiers of K Nearest Neighbor (KNN) and Decision Tree (DT) are used to evaluate the classification accuracy. At the final stage, the process of data classification is repeated as a number of different features on the parallel genetic algorithm. Number of features means the specifying the number of features selected from all features in the dataset. By use of such method, the number of desired features is obtained automatically by the algorithm. In most feature selection methods, the number of desired features is given to algorithm as the input. The number is obtained with regards to the issue and dimensions of dataset by try and error. In most high-dimension dataset, like data of intrusion detection system (NSL-KDD), it is difficult to find the desired number of features and numerous experiments are needed. This dataset includes 41 features. We consider the desired features in the range of 14-30 (17 cases). This interval is proposed for finding empirically the desired number of features. Note that, the repeated implementation of the algorithm with such number of features is time-consuming. For this purpose, we use the parallel implementation of the genetic algorithm for selecting the features with different number of features. When the final desired features are determined, the results of the intrusion detection system are shown on the best classification method (DT or KNN). The proposed algorithm is shown in figure 1. Figure 1 shows the flowchart of the proposed algorithm for the intrusion detection.

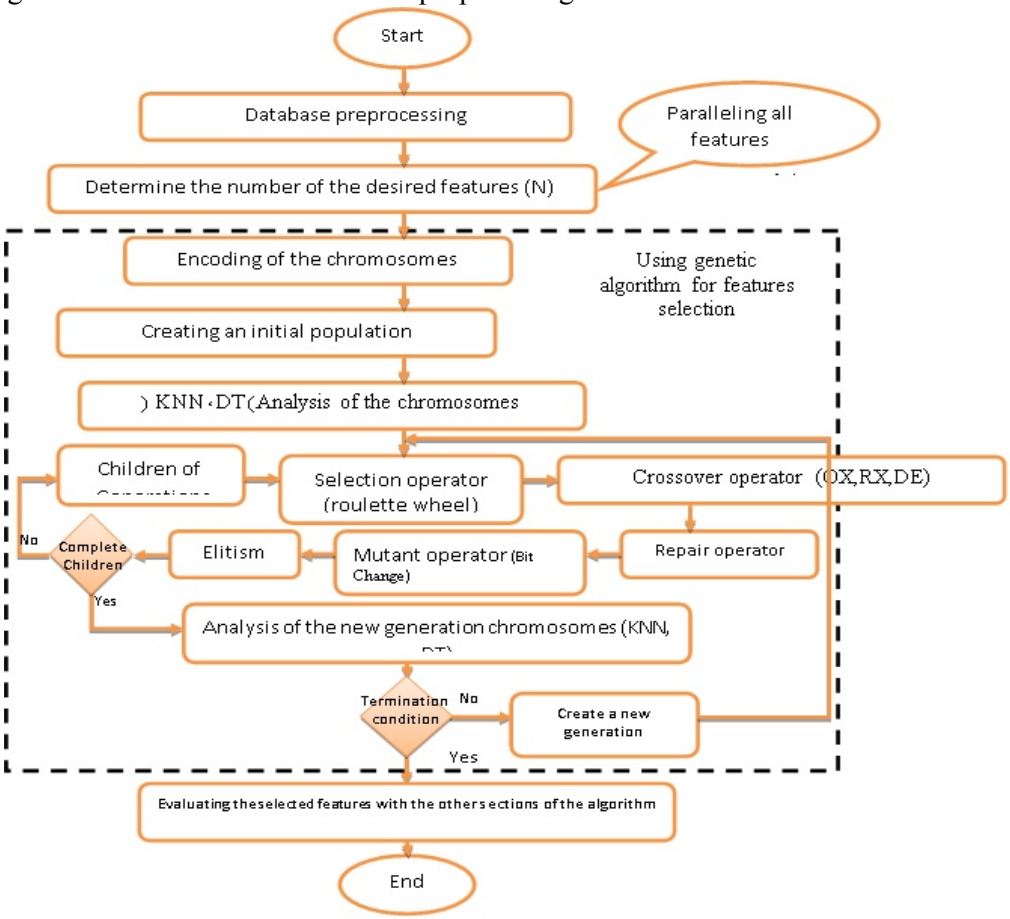



***Figure 1: The proposed algorithm for the feature selection***

### *3.1 Using genetic algorithm for features selection*

The schematic structure of the chromosomes is the first question that is posed in the mind that a person who is supposed to solve and optimize the issues by the genetic algorithm. In features selection, we take the number of each feature into account as a criterion for selecting that feature. In this case, the features which their numbers are on the chromosome will participate in data classification and others are not included. Chromosome length is determined in accordance with the desired number of features (DNF). The genetic algorithm starts with creating the primary population of chromosomes. In this study, we select randomly the chromosomes from the search space. NP parameter is used for displaying the number of population members. The fitness function is the error rate of data classification in features selection issue. Therefore, training samples are classified in terms of each chromosome (selected features) and the computed error is seen as its goodness criterion. Hence, our main issue is to minimize the objective function. In the present study, two well-known classifiers of KNN and DT are used for classifying data and computing the chromosomes' fitness criterion. The reason for the using these two classifiers is their high speed in large database classification. Since the successive calculation of the selected features' fitness is required in the feature selection method, priority is to use the methods which can classify data with high speed (compared to the accuracy). We have used several classification methods for classifying the training dataset NSL-KDD in the same conditions to select these two classifiers. Each method performs the classification operations with different accuracies in a special time interval. Results obtained from the used methods in Table 1 clearly show that the two classifiers of decision tree and k nearest neighbor are faster than the other methods. Hence, we use these two methods for classifying the data and computing the chromosomes' fitness criterion. Note that, the value of fitness for all chromosomes are calculated in parallel and for each chromosome, these two classifiers calculate the classification error concurrently on the data, and the minimum error is considered the fitness criterion for that chromosome.

***Table 1: Comparing different classification methods***

| Method | LVQ | KNN | DT | MLP | SVM |
|---|---|---|---|---|---|
| **Time(s)** | 1876.7 | **7.86** | **2.34** | 63.58 | 875.12 |
| **Accuracy (%)** | 99.71 | 99.04 | 99.32 | 88.62 | 99.61 |

In order to select the chromosomes for reproduction, Roulette Wheel is used. The most important operator in the genetic algorithm is the crossover operator. Pairs seen as the parent in the selection part, exchange their genes with each other and produce new members. There are different forms of crossover operator such as single-point, two-point, n-point, and uniform crossovers. In this study, the crossover operators are designed which will be discussed in the following.

**A. Ordered crossover (OX) operator**

This operator is much similar to the two-point crossover, but with this difference that it only produces a child. At first, two random situations between two genes are taken into account. Then, all genes between these two situations are copied from the first parent chromosome to the child chromosome. Finally, the other genes are copied from

the chromosome of the second parent to the child chromosome. The order of genes is not important in this issue, but it should be kept in mind that the repeated genes should not be observed in child chromosome. In figure 2, you see how such operator acts.

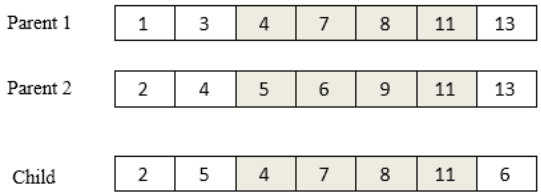



***Figure 2: Ordered crossover operator***

In this example, the positions 3 and 7 are the cutting points. The genes between these two positions are copied from the first parent (i.e. 4, 7, 8, and 11) to the child's chromosome. From the second parent, the 2nd, 5th and 6th genes are copied, since the 4th gene has already been added to the child's chromosome.

**B. Random crossover operator**

It is possible that there are features among the population chromosomes that are less repeated or have never been used. For this purpose, an operator is used which raise the probability of selecting the unobserved features. Random crossover operator produces the child chromosome randomly among the parents' genes and the other unobserved genes. This operator selects half of the child genes from two parents and the other half randomly. As seen in figure 3, this operator also produces one child.

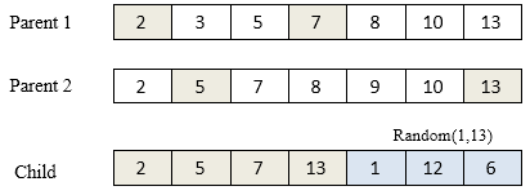



***Figure 3: Random crossover operator***

In this example, it is assumed that the total number of features is 13. Therefore, a value from 1 to 13 can be chosen for each gene. Here, the 2nd and 7th genes of the first parent and the 5th and 13th genes of the second parent were randomly selected and copied to the child's chromosome. Finally, the 1st, 6th and 12th features are randomly considered for the other genes.

**C. Differential evolution crossover operator**

This improved operator is the method given in research by [15]. In this operator, child is produced with regards to the differences between features of parent chromosomes. To converge the genetic algorithm faster, the best population chromosome is used. Our aim is to create a vector from experimental features. This vector produces a new vector in view of the difference between the weights of two members of population, i.e. $X_i^1$ and

$X_i^2$ and also, the best population member, i.e. $X_i^0$. The following relation shows how two selected feature vectors (parents) and the best population vector are combined to produce a mutated feature vector (child). The relation (1) shows that how two selective attribute vectors (parents) and the best population vector are combined to produce a Mutated feature vector (child) [15].

$$X_i^{new} = \begin{cases} X_i^0 & if\ rand(0,1) > C_r \\ X_i^0 + F \times (X_i^1 - X_i^2) & Otherwise \end{cases} \quad (1)$$

Where

F is a scale factor in [0, 1], the value of which controls the population evolution. $X_i^0$ is the Ith feature of the best population feature vector. Parameter i of each vector refers to the features implemented from 1 to DNF (Desired Number of Features). Parameter Cr shows the possibility of mutation for each feature. The proposed method searches the features within [1, NF]. It is possible that large differences are made between the values obtained in the search space while repeating such processes and the feature is selected that is not in permissible limit. Hence, the changes in the proposed operator are needed. Here, scale factor (F) is calculated dynamically as:

Specifically, the value of F produces the features with different values and determining the appropriate value for it plays an important role in the algorithm convergence. Hence, it is necessary to create a change in the proposed operator. In this study, the value of F is dynamically calculated according to the values of the characteristics Xi1 and Xi2 using the relation (2).

$$F = \frac{C_1 \times rand(0,1)}{max(X_i^1, X_i^2)} \quad (2)$$

In this relation, C1 is less than one. This method allows each population member to have fluctuations in special limits as per the feature size to reach the optimum solutions. As a result, this method contributes the improvement of the optimum solutions. But, based on the structure of this operator, the features are likely selected which are not within the permissible limit. Therefore, the repeated features should be replaced by new features. On the other hand, it is possible that by optimizing the real numbers, two repeated numbers are produced. Such status is not acceptable in features selection. For example, by choosing the features of the produced member as [32.53 28.14 65.86 32.72 12.86], the rounded features will be [32 28 66 32 13]. Since the feature 32 is repeated in twice, one feature should be replaced by a new feature. The distribution factor of the features is used to solve such problem (Haupt & Haupt, 2004). By using the distribution factor, repeated features and the out-of-limit ones are replaced by those features with the most distribution among the population. Here, we calculate the weights related to each feature in which probability of each feature is calculated as per its distribution factor. The distribution factor is calculated from fk feature (kth feature) as FDk.

It is calculated according to (3):

$$FD_k = a_1 \times \left(\frac{PD_k}{PD_k + ND_k}\right) + \frac{NF - D}{NF} \times \left(1 - \frac{(PD_k + ND_k)}{max_k(PD_k + ND_k)}\right) \quad (3)$$

Where,

PDk is the times that fk feature is used in good subsets (solutions). subset whose fitness value is less than the mean fitness of the total population. NDk is the times fk feature is used in bad subsets i.e. subset whose fitness value is more than the mean fitness of total population. NF is the total number of features, D is the desired number of the selected feature, and a1 is the positive constant for reflecting the importance of PD. In figure 4, an example of PDk and NDk is shown. Fitness criterion of the first four members is less than the mean fitness of the whole population. Therefore, PD is calculated based on such solutions. Distribution probability of PD for the first feature is 4/10, since there are four members in each feature. However, it is used only twice in other members (ND = 2/10)

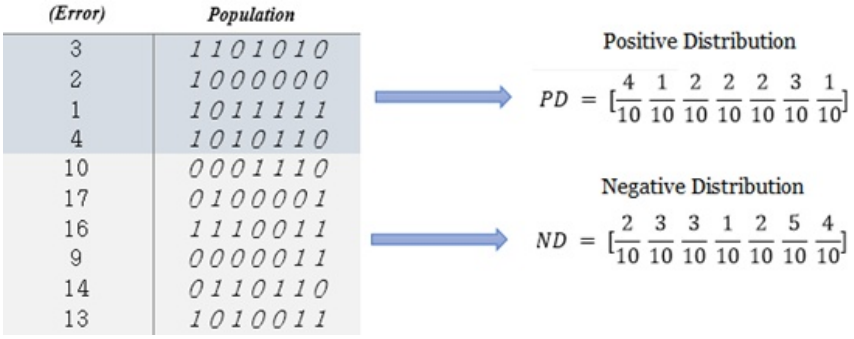



***Figure 4: An example of features distribution factor***

With regards to the previous example with a repeated feature, the aim is to correct the chromosome [32 28 66 32 13] in such a way that the repeated feature 32 is replaced by another appropriate feature. We suppose that the features rating is in accordance with the highest distribution factor as [78 34 21 19 68 74]. Therefore, feature 32 is replaced by feature 78 (highest distribution). As a result, chromosome is shown as [78 28 66 32 13].

### *3.2 Structure of Mutation genetic algorithm Structure mutation operator*

When the crossover is finished, the mutation operator is applied simultaneously to the three produced (child) chromosomes. Mutation is a random process in which the content of a gene is replaced by another gene to produce a new genetic structure. Role of mutation in the genetic algorithm is to recover the missed genetic materials or not found in the population. In this study, bit change mutation is used. This operator selects randomly a gene from a chromosome and then changes its content. For example, the first gene with a characteristic value of 2 can be changed from the child's chromosome of Figure 3 to the value of the 9. The result of this change will be applied if the fitness criterion is improved.

One of the interesting phenomena is the genetic algorithm in which the middle generations produce chromosomes very suitable in terms of value and goodness. Such chromosomes are likely destroyed due to the performance of crossover and mutation operators and are never reproduced. One way is to identify such cases and use of them in the next generations. This technique is called elitism which is practically effective in

finding the problem response. In fact, this method makes the good chromosomes to move directly toward the next generations. In this study, in each generation, only one chromosome is directly transferred to the next generation. This chromosome suggests the best solution in terms of classification. Since three proposed hybrid operators produce one child, they are repeated to the number of population members until the new chromosomes are produced. In the present research, the parallel method is used for producing the children. This means that the number of chromosomes required for production is n, n chromosomes are produced in parallel with regards to the selection operators. When the children are produced, current generation is directly replaced with the new generation and genetic cycle is repeated. The point which should be taken into account in designing any algorithm is the algorithm termination condition. In this study, a given number of generations are used for genetic algorithm termination condition. Due to constant implementation of algorithm for identifying the desired number of features automatically and also, for preventing the implementation of more generations, we consider fixed number for each part of algorithm.

### *3.3 Structure of parallel genetic algorithm*

Among the important capabilities of the genetic algorithm is the possibility of its implementation in parallel and also, searching the spaces which are too complicated or large. For example, calculating the people evaluation function can be done independent from the society and in parallel. The most important disadvantage of the intrusion detection methods is high computational cost and lack of guarantee for reaching the optimum response. High computational cost for this study can be resolved by implementing the algorithm in parallel on several computers or processors. Generally, it can be said that in issues where the search space is too complicated and large, it is less likely that the parallel genetic algorithms (PGA) are trapped in local minimums. As mentioned before, the desired number of features is determined automatically by algorithm. The number of different features to be examined by the algorithm is too high. Therefore, the constant implementation of the algorithm with such conditions leads to high computational costs. In this study, the parallel structure of the genetic algorithm is used for reducing the computational costs so that the several genetic algorithms are working in parallel with a different number of features. The genetic algorithm which cannot achieve the given classification accuracy after meeting the termination condition restarts with some other features. This process continues till finishing the list of the desired number of features.

In fact, instead of checking the number of features in sequence using the genetic algorithm, it is done in parallel and by assigning the number of attributes to each genetic algorithm. Finally, the number of features which has been achieved to the best classification accuracy is taken into account as the algorithm output.

## 4. Results and Experiments

In the present study, dataset NSL-KDD is used for evaluating the proposed method. To full access to this dataset, it can be used the following link [16].This dataset is an improved KDDCUP'99 dataset that is proposed to solve some of its inherent problems [17].The dataset of KDDCUP'99 is widely used to detect the dissonance. However, [17]

achieved two important issues in the statistical analysis of this dataset which were widely used to evaluate the performance of the systems which attenuate the results of the evaluation of the abnormal detection methods.

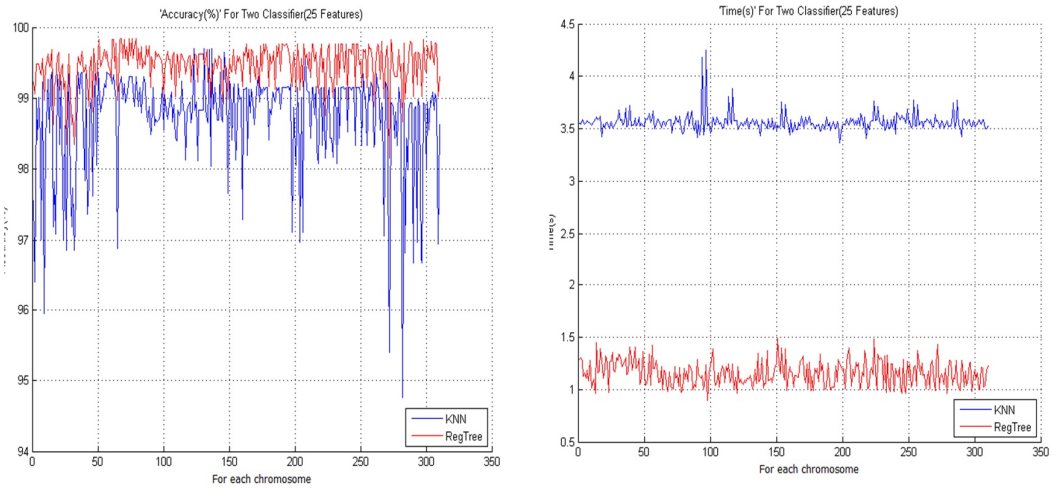
To solve these issues, they provided a new NSL-KDD dataset which consists of the records selected from the KDD dataset. NSL-KDD is presented in three sections of Percent Training Set, Train set and Test set with 25193, 125973 and 22543 samples, respectively. The dataset has consisted of 41 features and five classes that include a normal bunch and four types of Dos, R2L, U2R, and Probing attacks.

MATLAB version 2013 is used in this study for implementation and analysis of datasets. Matlab software is one the most widely used mathematical programming tools and has wide capabilities. MATLABPOOL capability in MATLAB allows user to use the parallel processing in computers with multi-core processors.

The results obtained from the experiments for enhancing the evaluation accuracy is the mean for 10 times test repetition. In implementation, population, number of generations, crossover rate and mutation rate are taken 170, 30, 0.95 and 0.15, respectively. Table 2 shows 25 features of dataset NSL-KDDCUP which is calculated automatically and by using the genetic algorithm. We have investigated the efficiency of two classifiers in terms of accuracy and speed on each produced chromosome and shown the results in table 5. With regards to the results of the feature selection algorithm, DT classification method is more efficient than KNN and for this purpose, the classification results are written based on this method. In table 3, accuracies of classes of DOS, Probe, U2R and R2L together with a normal class are given.

***Teble2: selected features of datasets NSL-KDD in the proposed method***

| Number Features | Name features | Number Features | Name features | Number Features | Name features |
|---|---|---|---|---|---|
| 2 | protocol_type | 14 | root_shell | 23 | Count |
| 3 | Service | 15 | su_attempted | 26 | srv_serror_rate |
| 4 | Flag | 16 | num_root | 27 | rerror_rate |
| 5 | src_bytes | 17 | num_file_creations | 29 | same_srv_rate |
| 6 | dst_bytes | 18 | num_shells | 30 | diff_srv_rate |
| 7 | Land | 19 | num_access_files | 34 | dst_host_same_srv_rate |
| 9 | Urgent | 20 | num_outbound_cmds | 38 | dst_host_serror_rate |
| 10 | Hot | 21 | is_host_login | 40 | dst_host_rerror_rate |
| 11 | num_failed_logins | | | | |




a) compare the accuracy for each chromosomes compare in terms of speed for each chromosomes

***Figure 5: comparison of classification KNN, DT in terms of speed and accuracy***

***Table 3: Percentage of proposed method detection by type of attack***

| Dataset | | Normal | DOS | Probe | U2R | R2L |
|---|---|---|---|---|---|---|
| **NSL-KDD** | Train | 99.98 | 100.0 | 99.99 | 90.38 | 99.50 |
| | Test | 99.91 | 99.83 | 99.88 | 94.50 | 99.35 |

For evaluating, the comparison must be made to the label assigned by the classification model to that attack. Various scenarios with FN, TN, FP, and TP values for a double-penetration detection system are shown in Table 4. This table is known as a confusion matrix.

***Table 4: Confusion matrix for intrusion detection data***

| Actual Records | Predicted Normal | Predicted Attack |
|---|---|---|
| **Normal** | TN | FP |
| **Intrusions (attacks)** | FN | TP |

The correct prediction of an intrusion detection system is determined by two TP and TN criteria. The most important criterion for determining the efficiency of a classification algorithm is the Accuracy criterion. This criterion calculates the total accuracy of a category. This criterion indicates that a few percent of the entire data set is properly categorized. Equation (4) shows how to calculate the correct criterion.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

Confusion matrix of intrusion detection system data for each class of attacks DOS, Probe, U2R, and R2L together with a normal class is calculated and shown in table 5. In this table, the number of records for each attack type together with the number of predictions is given.

***Table 5: Confusion matrix of NSL-KDD dataset by type of attack***

| **Actual Records** | | | **Predicted** | | | | |
|---|---|---|---|---|---|---|---|
| **Records Type** | **nsl-dataset** | **Number** | **Normal** | **DOS** | **Probe** | **U2R** | **R2L** |
| **Normal** | Train | **67343** | 67332 | 1 | 1 | 0 | 9 |
| | Test | **9710** | 9701 | 2 | 0 | 0 | 7 |
| **DOS** | Train | **45927** | 0 | 45926 | 1 | 0 | 0 |
| | Test | **7458** | 8 | 7445 | 3 | 0 | 2 |
| **Probe** | Train | **11656** | 1 | 0 | 11655 | 0 | 0 |
| | Test | **2421** | 3 | 0 | 2418 | 0 | 0 |
| **U2R** | Train | **52** | 4 | 0 | 0 | 47 | 1 |
| | Test | **200** | 7 | 0 | 2 | 189 | 2 |
| **R2L** | Train | **995** | 5 | 0 | 0 | 0 | 990 |
| | Test | **2754** | 13 | 0 | 0 | 5 | 2736 |

In what follows, to evaluate the aforementioned approach, the performance of the proposed system is compared with the other intrusion detection methods of which the experimental results on NSL-KDD are given. As can be seen from table 6, the proposed method has more accurate attacks compared to other intrusion detection systems for some attack and it also gives proper accuracy in other cases.

***Table 6: Comparison of proposed method with Other Intrusion detection methods by type of attack***

| **Classifier** | **Normal** | **DOS** | **Probe** | **U2R** | **R2L** | **Accuracy (%)** |
|---|---|---|---|---|---|---|
| SIPSO [18] | - | 99.80 | 99.70 | 97.50 | 82.50 | - |
| Adaptive IDS [4] | 66.51 | 88.64 | 99.15 | 66.51 | 20.88 | 75.15 |
| GA [7] | - | - | - | - | - | 68.51 |
| CSM [8] | - | - | - | - | - | **99.79** |
| RandomTree +IG [19] | 99.85 | **100.0** | 99.70 | 88.45 | 98.00 | 97.20 |
| Ant Colony [9] | 97.41 | 99.78 | 74.65 | 93.51 | 99.17 | 98.9 |
| MARS [10] | 99.71 | 99.97 | 99.85 | 76.00 | 98.75 | 92.75 |
| FV-GA [20] | 96.74 | 98.68 | 99.17 | **100.0** | **100.0** | 99.18 |
| **Proposed Method** | **99.91** | 99.83 | **99.88** | 94.50 | 99.35 | 99.76 |

## 5. Conclusion and Recommendations

In high-dimensional data, a high accuracy can be achieved due to using the appropriate methods for selecting features and necessary preprocessing steps. In this paper, we propose a method for selecting the proposed feature in which the number and set of the most suitable features are obtained using the genetic algorithm. In order to achieve better performance, a parallel implementation of the genetic algorithm has been used.

The parallel structure of the proposed genetic algorithm makes it possible to determine the number of features automatically without the need for information from the user by the algorithm. The results of the experiments show that the proposed method presented in this paper well categorized the penetration data and also, achieved an average accuracy of 99.76%. In general, the proposed method of this study is suitable for detecting a number of infiltrations in the KDD dataset, but some of the new attacks may have the normal behaviors, while their call parameters are abnormal. Consequently, such attacks cannot be detected by this method. For the future, it is suggested that besides using the above method, the input and output parameters in the system calling should also be used as the detect infiltration in order to detect more advanced attacks. Due to the sensitivity of time-sensitive intrusion detection, it is necessary to analyze the performance of the proposed method in terms of the duration of the detection of attacks in the form of an operational model.

## Reference

[1] Pan, Shengyi, Thomas Morris, and Uttam Adhikari. (2015). Developing a hybrid intrusion detection system using data mining for power systems. IEEE Transactions on Smart Grid 6.6: 3104-3113.

[2] Zuech, Richard, Taghi M. Khoshgoftaar, and Randall Wald. (2015). Intrusion detection and big heterogeneous data: a survey. Journal of Big Data 2.1: 1.

[3] Kenkre, Poonam Sinai, Anusha Pai, and Louella Colaco. (2015). Real time intrusion detection and prevention system. Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014. Springer International Publishing.

[4] Goyal, Anup, and Chetan Kumar, (2008). GA-NIDS: a genetic algorithm based network intrusion detection system. Northwestern university.

[5] Muda, Z., W. Yassin, M. N. Sulaiman, and N. I. Udzir, (2011). Intrusion detection based on K-Means clustering and Naïve Bayes classification. In Information Technology in Asia (CITA 11), 2011 7th International Conference on, pp. 1-6.

[6] Sriparna Saha, Ashok Singh Sairam, Asif Ekbal. (2012). Genetic Algorithm Combined with Support Vector Machine for Building an Intrusion Detection System, International Conference on Advances in Computing, Communications and Informatics (ICACCI-2012)

[7] Benaicha, Salah Eddine, Lalia Saoudi, Salah Eddine Bouhouita Guermeche, and Ouarda Lounis, (2014). Intrusion detection system using genetic algorithm. In Science and Information Conference (SAI), 2014, pp. 564-568.

[8] Chae, Hee-su, Byung-oh Jo, Sang-Hyun Choi, and Twaekyung Park, (2015). Feature Selection for Intrusion Detection using NSL-KDD. Recent Advances in Computer Science, ISBN: 978-960.

[9] Aghdam, Mehdi Hosseinzadeh, and Peyman Kabiri, (2016). Feature selection for intrusion detection system using ant colony optimization. International Journal of Network Security 18.3: 420-432.

[10] Mubarak, Shaik Liyakhat, (2016). Intrusion Detection System using SVM, SOM & NN.

[11] Kevric, J., Jukic, S., & Subasi, A. (2016). An effective combining classifier approach using tree algorithms for network intrusion detection. Neural Computing and Applications, 1-8.

[12] Mehdi Akbari," An Efficient Genetic Algorithm for Task Scheduling on Heterogeneous Computing Systems Based on TRIZ" Journal of Advances in Computer Research, Volume 9, Issue 3, Summer 2018, Page 103-132.

[13] Yaser Nemati, Pirooz Shamsinejad" Journal of Advances in Computer Research, Article 7, Volume 9, Issue 2 - Serial Number 32, Spring 2018, Page 103-112.

[14] Ali Nodehi," Determining Cluster-Heads in Mobile Ad-Hoc Networks Using Multi-Objective Evolutionary based Algorithm" Journal of Advances in Computer Research, Volume 9, Issue 3, Summer 2018, Page 133-151.

[15] Khushaba, Rami N., Ahmed Al-Ani, and Adel Al-Jumaily, (2011). Feature subset selection using differential evolution and a statistical repair mechanism. Expert Systems with Applications 38.9: 11515-11526.

[16] Nsl-kdd dataset for network based intrusion detection systems. Available on: http://nsl.cs.unb.ca/KDD/NSL-KDD.html, March 2009.

[17] Tavallaee M, Stakhanova N and Ghorbani AA, (2010). Towards credible evaluation of anomaly based intrusion detection methods, IEEE Transaction on System, Man and Cybernetics, Part-c, Applications and Reviews; 40(5):516-524.

[18] Warsi, Sana, Yogesh Rai, and Santosh Kushwaha, (2015). Selective Iteration based Particle Swarm Optimization (SIPSO) for Intrusion Detection System. International Journal of Computer Applications 124.17.

[19] Kohdayar, M., Asareh, A. & Aminilari, M. (2014). "Application of Machine Learning Hybrid Algorithm in Improving the Intrusion Detection Systems", National Conference on Computer Engineering and Information Technology Management, Tehran, Tolu Farzin Industry and Science Company.

[20] Bhattacharjee, P. S., Fujail, A. K. M., & Begum, S. A. (2017). Intrusion detection system for NSL-KDD data set using vectorised fitness function in genetic algorithm. Adv. Comput. Sci. Technol., 10(2), 235-246.