An Enhancement of the Replacement Steady State Genetic Algorithm for Intrusion Detection

Reyadh Naoum¹, Shatha Aziz², Firas Alabsi³

Abstract

In these days, Internet and computer systems face many intrusions, thus for this purpose we need to build a detection or prevention security system. Intrusion Detection System (IDS) is a system used to detect attacks, Steady State Genetic Algorithm (SSGA) is applied to support IDS by supplying the rule pool with additional data, these data can be used in testing phase to detect the attacks. The main goal of this research is to enhance Replacement steady state genetic algorithm to detect intrusions. This enhancement has been achieved by comparing replacement methods. This research proved that the Triple Tournament Replacement is better than Binary Tournament Replacement to increase Detection Rate and there are no effects on False Positive Rate. In this research represent the results of DR equal 100% for three types of attack (DoS, Probe and R2T), and 53% for U2R.

Keywords

Triple Tournament Replacement, Detection Rate, Intrusion Detection System, Steady State Genetic Algorithm.

1. Introduction

Intruder is defined as a system, program or person who tries to violate an information system or execute an illegal action [1]. Many studies and researches tried to build smart and strong security systems to protect computer systems from intrusions [2]. Security and trusted communication of information over internet and any other network is always under threat of intrusions. So, Intrusion Detection Systems have become a necessary component in terms of computer and network security. Genetic Algorithm (GA) is one of the most important approaches used to help Intrusion Detection issue. It aims to produce many solutions for the problem and select the best solution. Since there is no perfect solution to detect intrusions from violating system privacy, it is very important to be able to detect intrusions on time of occurrence and take actions to minimize the possible damage. Network Intrusion Detection System (NIDS) is one of the significant components of network security. NIDS must be efficient to increase the trust of using the system, by detecting intrusions with high Detection Rate (DR) and low False Positive Rate (FPR). The objectives of this research are Enhancing Steady State Genetic Algorithm (SSGA) by comparing replacement methods then choosing the best method to produce a new system for intrusion detection. The system can be evaluated as a good system if the enhancement affects the system by increasing DR and decreasing FPR. Finally, this research has compared the results that will be obtained with the earlier results produced from using Steady State Genetic Algorithm. The significance of this research is in supporting the existing IDSs. It deals with enhanced Steady State Genetic Algorithm with using new Replacement approach.

2. Theoretical Background

Intrusion Detection: Intrusion can be denoted as any set of actions that try to compromise the integrity, confidentiality or availability of a computer resource [1]. Intrusion Detection (ID) is a process of discovering the intrusion activity. It can be classified into two types Anomaly and Misuse Intrusion Detection [2]. Computer System Security can be defined as a process of protecting the main aspects for any computer system security. Those aspects are: Confidentiality, Integrity, and Availability, which are referred in the abbreviation CIA [3]. Intrusion Detection System (IDS) has become one of the hottest research areas in Computer Security. Attacks types are classified according to the following categories [4]:

1) Denial of Service Attack (DoS): Is an attack in which the attacker makes some...
computing or memory resources too busy or too full to prevent legitimate users access to system.

2) **User to Root Attack (U2R):** occurs when an attacker starts out with access to a normal user account on the system and is able to exploit some vulnerability to gain root access to the system.

3) **Remote to Local Attack (R2L):** occurs when an attacker who has the ability to send packets to a machine over a network, and then exploits the machine’s vulnerability to lawlessly gain local access as a user.

4) **Probing Attack:** is an attempt to learn information about a network of computers to look for exploits.

**Genetic Algorithm (GA):** Genetic Algorithm is “an adaptive heuristic search method based on population genetics”. Genetic Algorithm is a search method based on the principles of natural selection and genetics [5]. Genetic Algorithm is based on Darwin’s principles in optimizing the chromosome population of candidate solutions [6]. The algorithm starts with a population of chromosomes (individuals) which have a determined size, then evaluate fitness of chromosomes by measuring the value of fitness function for each one, then generate new generation by applying genetic operators such as Selection, Crossover and Mutation frequently, until the stopping criterion achieved and the best chromosomes presented.

**Steady State Genetic Algorithm (SSGA) elements**

**Population:** Population is a set of individuals (chromosomes) of a specified size. The size of population is determined according to the nature of the problem. Generally, the population is generated randomly [5].

**Evaluation:** Each individual has a fitness value used to evaluate the fitness for each chromosome; fitness value evaluates the quality of each chromosome so the high fitness value gives chromosomes with high probability to be selected in the Selection stage [7]. A Reward Penalty based Fitness Function presented in [8] as the following equation:

\[ Fitness = 2 + \frac{AB - A}{AB + A} + \frac{AB}{X} - \frac{A}{Y} \]  

(1)

The values of A and AB depend on the condition and action parts. X is the maximum value of AB in the population. Y is the maximum value of A in the population.

**Encoding:** Encoding is one of the significant processes in Genetic Algorithm to represent solutions. The gene in Genetic Algorithm is considered the problem parameter which can be encoded as one of encoding methods. There are different methods for encoding such as: Binary encoding, Integer encoding and real encoding.

**Selection:** Selection is processes of choosing individuals (parents) from current population to implement operations of Crossover and Mutation on them to generate new individuals [7]. There are different methods of Selection such as Roulette wheel Selection, Elitist Selection, Ranking Selection, Stochastic Universal Sampling and Tournament Selection.

**Replacement:** It is a process performed on the worst individuals to be replaced by better new individuals. There are two methods of Replacement:

- **Binary Tournament Replacement (BTR)**
  It will choose the best chromosome from two chromosome according to their fitness values.

  For \( n = \{1,2\} \), random numbers \( i,j \in \{1,2,\ldots,N\} \), \( i \neq j \)

  Where:
  
  Replace\( (n) \): individual \( n \) that will be replaced
  
  \( F(\text{ind}(i)) \): fitness of individual \( i \) 
  
  \( F(\text{ind}(j)) \): fitness of individual \( j \) 

- **Triple Tournament Replacement (TTR)**
  It will replace the worst two chromosomes between three chromosomes by the chromosome with the best fitness value.

  For \( n = \{1,2,3\} \), random numbers \( i,j,k \in \{1,2,\ldots,N\} \), \( i \neq j, i \neq k, j \neq k \)

  Where:
  
  Replace\( (n) \): individual \( n \) that will be replaced
  
  \( F(\text{ind}(i)) \): fitness of individual \( i \) 
  
  \( F(\text{ind}(j)) \): fitness of individual \( j \) 
  
  \( F(\text{ind}(k)) \): fitness of individual \( k \)
Stopping criteria: There is a repetition in the evolution process of GA, until satisfaction of the stopping condition. This research stops the evolution when there aren’t new chromosomes to be created through additional generations; e.g. suppose that first generation created 200 of chromosomes, at the next generations the production will be decreased, and so on until the production of new chromosomes approach to Zero. Thereafter the evolution process will be stopped.

KDD Cup 99 Dataset: In 1998, the United States Defence Advanced Research Projects Agency (DARPA) funded an “Intrusion Detection Evaluation Program (IDEP)” managed by the Lincoln Laboratory at the Massachusetts Institute of Technology. This program was achieved to build a data set that would help to evaluate different intrusion detection systems (IDS) in order to evaluate their strengths and weaknesses. This data set is popularly known as DARPA 1998 data set [9][10].

3. Literature Review & Related Works

In 2000, Reeves [13] presented study about GA and explained the detailed steps are conducted in the algorithm to obtain the best solution. The researchers claimed that GA aims to produce many solutions to solve specific problem such as the problem of “salesman” which is familiar in Artificial Intelligence science. The researchers also studied insight in the GA procedure to indicate how this algorithm can be exploited to solve any problem type such as: finance, medical, mathematical and technical problems. In 2004, Haupt and Haupt [14] explained practical GA and indicated the concept of finding best solution by optimizing GA. The author explained that GA can be used for many purposes and objectives in solving different problems. Also, the study explained the encoding and decoding processes that are used in GA to select the best chromosomes. In 2012, Al-Sabbah [15] presented a model for enhancing the colour image using the Steady State Genetic Algorithm (SSGA), and used modified fitness function to achieve more accurate result with less noise. The main objective of this research is to process the image so that the result is more suitable than the original image. The advantage of this research is combining the chromaity components of the image.

In 2009, Al-Sharafat [16] highlighted the techniques used for detecting intrusion using Genetic Algorithm (GA). Author represented the intrusion that attacks the network and the security must be placed on specific situations on the network. However, the author used Steady State Genetic Algorithm (SSGA) that gives...
Detection Rate (DR) reaches to (97.45)%. The author also compares between the algorithms that usually used for detecting intrusion using DR parameter. In 2012, Hoque et al [1] proposed detection system using GA and applying their algorithm on specific datasets of intrusions detection (KDD’99) which are based on 1998 DARPA. Their implementation is based on defining datasets as chromosomes and then making GA produces the parents and children to select the best chromosome. However, they chose range value about 0.125 for Selection purpose and they conducted Crossover and Mutation in the next stages. They got better Detection Rate for Denial of Service and User-to-Root besides to close detection for Probe and Remote-to-Local.

In 2012, Naoum et al [17] classified intrusions by using an enhanced Resilience Back Propagation Neural Network. After they completed their system and tested the proposed system, they got 94.7% average Detection Rate, and 15.7% False Positive Rate. In 2012, Kshiragar et al [18] used GA with Data Mining technique to develop Intrusion Detection System. The proposed system depended on data mining methods and they proposed their system in model called “Data Mining Hybrid IDS”. The model contains three sensors located at network sensor manager that is connected with data “warehouse unit”. The data is transferred between warehouse and network sensors and pattern mining. The proposed system enables administrator to prevent intruder activity by Data Mining techniques and sensor snort located on the network. In 2013, Mostaque [19] designed an Intrusion Detection System (IDS), by applying Genetic Algorithm (GA) and Fuzzy Logic to efficiently detect different types of attacks within a network. The proposed Fuzzy Logic-based System could be able to detect the intrusive activities of the computer networks as the rule base holds a better set of rules. The data used of the proposed Intrusion Detection System are received from the KDD Cup 99 Intrusion Detection benchmark dataset. The experimental results illustrate that the proposed system get higher accuracy in determining whether the records are normal or abnormal and obtained good detection rate.

4. Proposed Model & Methodology

This research deals with the enhanced Steady State Genetic Algorithm (SSGA) for Intrusion Detection (ID), through applying the best method of Replacement by comparing Replacement methods: Binary Tournament Replacement , Triple Tournament Replacement, and choose the best one hopping to get the new results. The following figure displays our own proposed model:

Figure 1: Proposed Model

The structure of the proposed IDS is described as the following:

Environment:
The KDD CUP 99 datasets will be the environment for implementation of the proposed model. In this research 10% of original dataset will be used as a sample for training and testing.

Training Dataset:
From training dataset, the system will start by receiving data to produce new rules from received record. This research used 9% as training dataset, which is approximately 444618 records.

Testing Dataset:
A part of the dataset can be used for examining the matching between packet and the rules stored in the rule pool, this research used 1% as testing dataset, which approximately 49403 records.

Classifier: Classifier used for classifying the data; in this phase the message that comes from training dataset will be received and then represent the message in a real representation to format the rule as a chromosome. The values of the most significant features vary from binary to real numbers. Also, the classifier will determine the condition and the action part of the rule. The condition part of the rule has a combination of values related to set of features. The relation between the condition values and the type of attack represents the action part. The last part of classifier is to filter the message from redundancy and deal with the most significant features.
Distinct Rules: This database will be produced after the following operations:
- Classifying data to the four classes (DoS, Probe, R2L and U2R).
- Creating the rules as Condition-Action form.
- Filtering the rules with the best significant features.
- Removing the redundant rules from the rules dataset.

Steady State Genetic Algorithm Unit: In this unit, Genetic Algorithm will be used to generate new rules from the existing data.

Chromosomes Evaluation (First Phase): The chromosomes will be evaluated by computing fitness value for each chromosome in order to be selected in the next stage.

This research will use Reward Penalty based Fitness Function proposed by [8].

- Selection: Selecting the appropriate individual can be done by using Elitist Selection. Elitist Selection gets the best results when it is used together with Uniform Crossover within Steady State Genetic Algorithm, depending on [20].
- Crossover: At this stage, Uniform Crossover will be used, where genes are randomly exchanged at random points within a chromosome to produce two new offsprings.
- Mutation: Mutation will use Flip Bit method If the feature has discrete value, by flipping the value of a gene that is chosen randomly. Otherwise, if the feature's value is continuous, the new value will be equal to a random number of specific ranges.

- Chromosomes Evaluation (Second Phase): This stage will be used to evaluate the generated chromosomes by using Reward Penalty based Fitness Function. Evaluating chromosomes at this stage helps in applying Replacement stage.

- Replacement: Apply Replacement by comparing the Replacement methods, and then decide which one gives the best result.

- Check the Stopping Criteria: Checking the stopping criteria can be checked if there are no additional new rules to be produced, then the Genetic Algorithm will be stopped; otherwise, the Genetic Algorithm creates an additional generation.

Rules Pool: It will contain the rules that are collected from training data and Steady State Genetic Algorithm unit in order to be used in the testing phase.

Testing (Matching): In this phase, the proposed system will try to match the received data with the existent rules in the rules pool in order to recognize the data and detect the intrusions. If it satisfied the matching condition, then the Alarm of Intrusion Detection will be appeared, otherwise, it is a Normal behaviour.

System Evaluation: Evaluating the proposed system will be done by calculating Detection Rate (DR) and False Positive Rate (FPR) then comparing the results with others.

The DR and FPR were calculated using the following calculations:

\[
DR = \frac{TP}{TP + TN} \quad (2)
\]

\[
FPR = \frac{FP}{TN + FP} \quad (3)
\]

Where TP refers to True Positive, TN refers to True Negative and FP refers to False Positive.

5. Experimental Results

In this research, we present the experimental results through the execution of proposed Intrusion Detection System which has been supported by Steady State Genetic Algorithm.

The Results of Detection Rate (DR) and False Positive Rate (FPR): As a result of this research, Intrusion Detection System has been built and supported by Steady State Genetic Algorithm. The goal was to get high DR, and to get low FPR. After system execution, we got the following results of DR and FPR with Binary Replacement for each type of attack:

| Table 1: Results of DR and FPR using Binary Replacement |
|------------------|----------------|--------|--------|
| DoS | Probe | U2R | R2L |
| DR  | 97% | 100% | 40% | 100% |
| FPR | 0.0232 | 1.84 | 1.866 | 2.195 |

And the results of DR and FPR with Triple Replacement for each attack type are:

| Table 2: Results of DR and FPR using triple Replacement |
|------------------|----------------|--------|--------|
| DoS | Probe | U2R | R2L |
| DR  | 100% | 100% | 53% | 100% |
| FPR | 0.0232 | 1.84 | 1.866 | 2.195 |

Thus we conclude that:
- Both of Binary and Triple Replacement have the same values of FPR, but they are different in the value of DR.
- The increasing in the number of generations leads to increase DR, but the values of FPR are not affected by the number of generations.

491
-Triple Replacement method produced more accurate results in DR than Binary Replacement.

Comparing research results with others: The first comparing procedure is with [21] which used Binary replacement in Genetic Algorithm with Misuse Intrusion Detection System. But [21] it has been used 5% as a sample of training dataset while this research used 9% as a sample of training dataset. The criterion of comparing is the DR of each attack.

Table 3: First comparison

<table>
<thead>
<tr>
<th></th>
<th>DoS</th>
<th>Probe</th>
<th>U2R</th>
<th>R2L</th>
</tr>
</thead>
<tbody>
<tr>
<td>This research</td>
<td>100%</td>
<td>100%</td>
<td>53%</td>
<td>100%</td>
</tr>
<tr>
<td>Alabsi</td>
<td>94%</td>
<td>100%</td>
<td>86%</td>
<td>100%</td>
</tr>
</tbody>
</table>

The second comparing procedure will be between our results with Stewart and Alsharafat results [22][16]. The criteria of this comparison will be the average of DR and FPR. Stewart [22] used hybrid system of Genetic Algorithm with Neural Networks for Intrusion Detection System. He found that the average of DR is equal to 79.67% which is less than this research results. Alsharafat [16] used anomaly detection with her IDS and she got the average of DR equal to 98.9% which is greater than the results of this research. Stewart [22] got the average FPR is 2.69%, which is worse than the results of our research. Al-sharafat got the average of FPR equal 0.094% which is more accurate than the results of this research. The following table shows these results:

Table 4: Second comparison

<table>
<thead>
<tr>
<th></th>
<th>Average of DR</th>
<th>Average of FPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our result</td>
<td>88.25%</td>
<td>1.48%</td>
</tr>
<tr>
<td>Alsharafat</td>
<td>98.9%</td>
<td>0.094%</td>
</tr>
<tr>
<td>Stewart</td>
<td>79.67%</td>
<td>2.69%</td>
</tr>
</tbody>
</table>

6. Conclusion

In this research, an Intrusion Detection System was built and we focus on enhancing the Replacement of Steady State Genetic Algorithm to increase Detection Rate. The environment of this research was the KDD Cup 99 dataset, all experiments and evaluations are performed by using 10% of the whole dataset. The inputs to the system are two subsets: training dataset which is 9% and testing dataset which is 1% of KDD Cup 99 dataset. We conclude from this research that Triple Replacement produces more accurate results than Binary Replacement, according to the value of DR, number of generations and number of new chromosomes. Also, we found that Triple Replacement produces DR equals to 100% for the following attack types (DoS, Probe, and R2L), but U2R attack produces a result of DR equals 53% which still has better result than U2R in Binary Replacement which produces DR equals 40%. This research improves the Fitness Function to be fit for finding a Fitness Value that is proper for new chromosomes that produced from GA.

References


2. Foundations of Mathematics, Volume two, Pub. at press of Basrah University, 1982
5. 116 other publications.

Shatha Azez: She has B.Sc (Mathematics) from Baghdad University, Iraq (2001). She is currently a M.Sc. student at the Department of Computer Information System in Middle East University. Her current research interests in developing GA and its applications.

Firas AlAbsi: He has B.Sc (Computer Science), AlZaytoonah University, Jordan (2005) and M.Sc. (Computer Science), Middle East University MEU, Jordan (2012). He interested in Evolutionary Algorithms and Optimization.