Fitness Function for Genetic Algorithm used in Intrusion Detection System

Firas Alabsi
College of Information Technology
Middle East University
Amman, Jordan

Reyadh Naoum
College of Information Technology
Middle East University
Amman, Jordan

Abstract

Computer network usage increased rapidly at the last decades, the intruders tried to satisfy their needs by many types of attack depending on the intruder objectives, this encourage the researchers to find more and more solutions to detect those attacks. Intrusion Detection System used to detect the attack. Genetic Algorithm used to support IDS. Fitness Function is helpful in chromosome evaluation which is a Genetic Algorithm part. The problem is to find a suitable Fitness Function for a chromosome evaluation to get a solution for Intrusion Detection. This paper suggests a new reasonable Fitness Function using Reward-Penalty technique to evaluate population chromosomes efficiently. This technique used to give reward to the good chromosome and to apply a penalty on the bad chromosome. This paper will show the Fitness Function, discuss it and compare it with another Fitness Function to check its validity.

Key words: Intrusion Detection, Intrusion Detection System, Genetic Algorithm, Fitness Function.

1. Introduction

Intrusion Detection (ID) is defined by [1] as: "the process of monitoring the events occurring in a computer system or network, and analyzing them for signs of possible incidents, which are violations or imminent threats of violation of computer security polices, acceptable use policies, or standard security practices". Intrusion Detection System (IDS) is an application implements the process of Intrusion Detection. If the system detects intrusions within network, then the system is called Network IDS.

Network IDS can be classified into two types according to the detecting approach, the first type is detecting the intrusions using known patterns, by comparing received pattern with the existence known patterns; this approach is called misuse intrusion detection system, but if the intrusion detection is done by measuring the deviation from the normal baseline, this approach is called anomaly intrusion detection. By the way, the results of this paper deals with Network Intrusion Detection System under Misuse Analysis.

Genetic Algorithm is an evolutionary algorithm for search and optimization, it mimics the genetic operations in the human body, and it is used recently to support intrusion detection systems, by creating new rule from available rules. This technique makes a great chance to detect the intrusions by the systems customized for this job.

Genetic Algorithm contains of a sequence of operations, which are: Selection, Crossover, Mutation and sometimes Replacement, but the first operation is depending on the fitness value that obtained by Fitness Function.

2. Related Work

[6] Presented Genetic Algorithm to identify the attack type of connection, the algorithm used different features in network connections to generate a classification rule set; they used the fitness function given by the formula

$$F = \frac{a}{A} - \frac{b}{B} \quad (1)$$

Where:
A: Total of attacks.
a: Number of attack connections the individual correctly classified.
B: Normal connections in the population.
b: number of normal connections a network correctly classified.
They set a threshold value of 0.95; they select the individual which have a fitness value > 0.95.

[7] Used GA to detect Denial of Service (DOS) and Probe type of attacks, they used a fitness function:

$$Fitness = \frac{f(x)}{f(Sum)} \quad (2)$$

Where f(x) is the fitness of entity x, and f(sum) is the total fitness of all entities.

[8] Used Genetic Algorithm for Intrusion Detection System, he calculated the fitness function by calculate the following four equations:

$$Outcome = \sum_{i=1}^{57} Matched * Weight(i) \quad (3)$$

$$\Delta = |Outcome - SuspiciousLevel| \quad (4)$$

$$Penality = \frac{\Delta * Ranking}{100} \quad (5)$$

$$Fitness = 1 - Penality \quad (6)$$

Using equation (3) the outcome is calculated based on whether the A field of connection matched the pre-classified data set and then multiply the weight of that field, the value of matched is 0 or 1.
In the equation (4), the actual value of suspicious Level reflects observations from historical data.
In the equation (5), ranking indicates whether or not the intrusion is easy to identify.
Finally the value of fitness computed in equation (6) using the penalty.

3. Problem Statement

Intrusion Detection System used to protect the system against malicious activities. Genetic Algorithms applied to support Intrusion Detection Systems. Genetic Algorithm can't be done without selection process which depends mainly on fitness value that obtained using Fitness Function.

But, chromosomes vary in their strength and weakness. Hence, Fitness Function must take two points in its consideration; first point: the reward must be as more as the chromosome strength, second point: the penalty must be as more as the chromosome weakness. This can be done using the suggested reward-penalty based Fitness Function.

4. Reward-Penalty based Fitness Function

Using vb.net 2010 and SQL server 2008, the data of 5% of KDDCup99 classified into 5 main categories; Normal, Dos, Probe, U2R and R2L. Each category record was compared to the whole data. Actually there are little features to be compared rather than comparing 41 features. According to [5], just 5 features used in the comparison in each separated category.

After classification stage, there are 5 tables, each table for just one category, each table has got a name as the category type and included 8 columns as the following: id Column, 5 columns to have a 5 features, and another two columns; A, and AB.

To understand the reason of creating column A, and column AB, suppose there are 5 features for DOS category, each feature value should be in a specific range or equal to specific value in order to evaluate the record as DOS, but in such cases, the five features get the same values as a record in DOS but still not DOS because of specific value of one or more of the hidden features.

Suppose that the features' values are a condition part and the category name is an action, then for each record compared with the whole 5% of CDDCup99, if the condition and action of the selected record equal to the condition and action of the Compared record, then this will increase the value of column AB of the selected record by 1. Else if the condition of the selected record equal to the condition of the compared record but the actions of both records don't meet each other, then the value of column A of the selected record will increased by one.

The new fitness function will depend mainly on the values of A and AB, the formula of the function is as the following:

$$Fitness = 2 + \frac{AB - A}{AB + A} + \frac{AB}{X} - \frac{A}{Y} \qquad (7)$$

Where:
X = the maximum value of AB in the population.
Y = the maximum value of A in the population.

Now, let us discuss the content of the function:
(AB/(AB+A)) gives the rate of the AB value in proportion to the sum of AB and A values, the resulted value will reflect the strength of the record.

(A/(AB+A)) gives the rate of the A value in proportion to the sum of AB and A values, the resulted value will reflect the weakness of the record.

To obtain the importance and strength of the record, one can subtract the weakness value from the strength value by calculating ((AB-A)/(AB+A)) as in the function above.

Now, stop for a moment and suppose that there are two records with the following values:

Table 1: Similar Fitness Values

Record	A	AB	Fitness = ((AB-A)/(AB+A))
Rec1	0	1	1
Rec2	0	5	1

But, in such cases the resulted value will not be accurate because it will deal with record1 and record2 as the same strength whereas it is clear that record2 is stronger than record1 because of the value of AB, so the function should be supported with other positive and negative values to apply policy of reward and penalty to the records as following:

AB/X: gives the rate which reflects the strength of the record depending on the most strongest record in the population, the resulted value will be equal to Zero in the worst case (If AB value = 0) and will be equal to One in the best case if the AB value of that record is the highest AB value in the population, so it is logically should be added to the function to reward the record.

A/Y: gives the rate which reflects the weakness of the record depending on the most weakness record in the population, the resulted value will be equal to Zero in the best case (If A value =0) and will be equal to One in the worst case if the A value of that record is the highest value in the population, so the value of A/Y must be subtracted from the function to give the penalty on the record.

Now, assume that the record with best case, so AB value of that record is the highest AB value in the AB column, and A value is equal to Zero, this means that Fitness = 2, in other hand, assume that the record with worst case, so A value of that record is the highest A value in the A column, and AB value is equal to Zero, this means that fitness = -2 , but the fitness value provided by the fitness function must assign a non-negative cost to each candidate [2], so the constant value of 2 will be added to the function to get fitness value equal to 0 in the worst case, and fitness value equal to 4 in the best case, in this manner, the fitness value will be positive and in the interval [0,4] at any case.

5. Experiments and Results

Using Vb.Net 2010 and SQL server 2008, the system has been built to calculate A value, AB value and Fitness value for each record in the attacks tables.

5.1 Fitness Results

The following table is from the real data set, the table contents of column A and column AB are filled according to the comparison process described above with a simple population of 4 records for each category, whereas the contents of the column Fitness is calculated using suggested fitness function.

Table 2: Real Data

Dos			**Normal**		
A	**AB**	**Fitness**	**A**	**AB**	**Fitness**
3	**419**	**3.985**	**0**	**1**	**3.143**
5	**280**	**3.632**	**0**	**3**	**3.429**
5687	**18**	**0.049**	**50**	**7**	**1.246**
23	**5**	**1.365**	**44**	**4**	**0.858**
R2L			**Probe**		
A	**AB**	**Fitness**	**A**	**AB**	**Fitness**
180930	**11**	**0.016**	**130691**	**2**	**0.002**
2114	**714**	**2.493**	**242**	**12**	**1.107**
0	**4**	**3.006**	**0**	**856**	**4.000**
0	**16**	**3.022**	**0**	**6**	**3.007**
			U2R		
			A	**AB**	**Fitness**
			134917	**6**	**1.000**
			1	**4**	**3.267**
			818	**3**	**1.501**
			10	**1**	**1.348**

5.2 Discussion of Fitness Results

Now, Observe that normal record with AB = 3 is more fit than normal record with A=1 in the case of A=0 in both records.

Observe that DOS record with AB = 5 is more fit than DOS record with AB = 18 because the first record has less A value than the other.

Observe the best case Probe record with fitness value = 4 that mean constant number (2) + 1 (because A value = 0) + 1 (Because AB is the greatest AB value in the population).

Observe the R2L record with AB = 4 is more fit than the R2L record with AB = 714 because of the high value of A for the record with AB = 714

Observe that U2R record with A = 818 and AB = 3 is more fit than U2R record with A=10 and AB = 1, because the maximum value of A is very high, and the maximum value of AB is very low, in these cases the reward and penalty issue affect the fitness value obviously.

5.3 Comparing Strategy

In order to approve the validity of the new fitness function, another fitness function should be tested to get the results and compare the results with new fitness function results.

If the fitness value of the rule X is greater than fitness value of the rule Y according to the first fitness function, then the fitness value of the rule X also greater than fitness value of the rule Y according to the second fitness function. For any record in the population there are two results R1 and R2 as the following two equations:

$$R1 = \frac{FitnessValue1}{MaxFitnessValue1inPopulation} \quad (8)$$

$$R2 = \frac{FitnessValue2}{MaxFitnessValue2inPopulation} \quad (9)$$

Where: Fitness Value 1 is the result of the reward-penalty based Fitness Function, and Fitness Value 2 is the result of the second Fitness Function of the same record. To say that the new fitness function is getting a good result, the values of R1 and R2 must be closed to each other.

Some of the researches [3][4] used Support Confidence Framework as a fitness function, the following equation used as fitness function:

$$FitnessFunction = t1 * Support + t2 * Confidence \quad (10)$$

Where:
Support: indicates the recurrence of AB within all rules in the population.
Confidence: indicates the recurrence of AB within all rules that have the same condition.
t1and t2 were used as thresholds to balance between support value and confidence value, assume that (t1 = 0.0257) and (t2 = 0.9843).

To get the accurate results, for each record in the population, we calculate fitness value 1 using Fitness Function 1 and fitness value 2 using Fitness Function 2, the second step is to find the values of R1 and R2 using the equations 8 and 9, the third step is to find the result of the following equation:

$$R3 = \frac{\sum_{i=1}^{N} R1 - R2}{N} \quad (11)$$

Where,
N: the number of records in the population.
To judge that both Fitness Functions getting the same results in assigning the appropriate fitness value to each record in the population, the result of R3 must approach to zero.

5.4 Comparing Results

Using Vb.Net and SQL server, the system has been built for a population of 68 records of R2L attack. For each record in the population the system calculated the values of A, AB, Fitness Value 1, Fitness Value 2, R1, And R2. Fitness Value 1 and R1 are related to the Reward-Penalty based Fitness Function, whereas Fitness Value 2 and R2 are related to the Support-Confidence Framework Fitness Function.
The following table contains some of the records and their values:

- **Table 3: Real Data from the Comparison System**

A	AB	Fit. Val.1	Fit. Val 2	R1	R2
180930	11	0.016	0.004	0.005	0.004
0	1	3.001	0.985	0.953	0.961
0	51	3.071	1.004	0.975	0.979
0	2	3.003	0.985	0.953	0.961
0	4	3.006	0.986	0.954	0.962
0	3	3.004	0.985	0.954	0.962
0	4	3.006	0.986	0.954	0.962
0	6	3.008	0.987	0.955	0.963
0	16	3.022	0.990	0.959	0.966
0	37	3.052	0.998	0.969	0.974
0	91	3.127	1.019	0.993	0.994
0	107	3.15	1.025	1.000	1.000

The results showed that R1 and R2 are closed to each other, finally the result of R3 was calculated using calculation (11) and it is approach to Zero, (R3 = - 0.0001).

6. Conclusion

This paper represented a new Fitness Function that determine the fitness value according to the condition-action recurrence, condition-only recurrence, and reward-penalty technique, the paper also described the function in details with discussion of best and worst case, the paper represented a table with a real value and discussed many cases in order to prove that new Fitness Function is working properly and effectively to help in reasonable evaluation, finally there are a comparison between Reward-Penalty based Fitness Function and another functions, the result show us that new Fitness Function is able to get a good results in using Genetic Algorithm for Misuse Network Intrusion Detection Systems.

References

Scarfone, K. Mell, P. (2007). Guide to intrusion detection and prevention systems (IDPS). *National Institute of Standards and Technology*. Special publication 800-94, Page 2-1 from: http://csrc.nist.gov/publications/nistpubs/800-94/SP800-94.pdf

Bottaci,L.,2001, A Genetic Algorithm Fitness Function for mutation testing presented at SEMINAL 2001,International workshop on software engineering using Metaheuristic Innovative algorithm, a workshop at 23-rd Int. Conference on Software Engineering, Toronto, May 12-19. From: http://www2.hull.ac.uk/science/pdf/workshop.pdf

Selvakani,S. Rajesh,R.S.(2007). Genetic algorithm for framing rules for intrusion detection. International Journal for Computer Science and Network Security IJCSNS, VOL(7), No(11), 285-290. from: http://paper.ijcsns.org/07_book/200711/20071144.pdf

Berlanga.F.J., Del Jesus.M.J., Gatco.M.J., Herrera.F.,A Genetic-Programming-Based Approach for the Learning of Compact Fuzzy Rule-Based Classification Systems, the eighth International Conference on Artificial Intelligence and Soft Computing (ICAISC), Zakopane, Poland, on 25-29 June 2006, PP 182-191. From: http://sci2s.ugr.es/docencia/doctoMineriaDatos/Ber06-ICAIS.pdf

Mukkamala, S., Sung, A., Abrham, A., (2004), 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, From: http://www.rmltech.com/doclink/LGP%20Based%20IDS.pdf

Goyal, A., Kumar, C., A Genetic Algorithm based Network Intrusion Detection System, not published Electrical Engineering and Computer Science Northwestern University Evanston(2007). From: http://www.cs.northwestern.edu/~ago210/ganids/GANIDS.pdf

Uppalaiah, B., Anand, K., Narsimha, B., Swaraj, S., Bharat, T., Genetic Algorithm Approach to Intrusion Detection System. International Journal of Computer Science and Technology IJCST. VOL 3, Issue 1, March 2012. From: http://www.ijcst.com/vol31/1/uppaliah.pdf

Li, W., Using Genetic Algorithm for Network Intrusion Detection, (2004), In proceeding of the United States Department of Energy Cyber Security Group 2004 Training Conference. From: http://www.security.cse.msstate.edu/docs/Publications/wli/DOECSG2004.pdf