Epigenetic Algorithm for Performing Intrusion Detection System

Mehdi EZZARII,
Institut National des Postes et Télécomunications (INPT),
Rabat, Morocco
ezzarii.mehdi@gmail.com

Hamid ELGHAZI,
Institut National des Postes et Télécomunications (INPT)
Rabat, Morocco
h.elghazi@inpt.ac.ma

Hassan EL GHAZI,
Institut National des Postes et Télécomunications (INPT),
Rabat, Morocco
elghazi@inpt.ac.ma

Tayeb SADIKI,
Université International de Rabat (UIR)
Rabat, Morocco
tayeb.sadiki@uir.ac.ma

***Abstract*—Intrusion Detection System is one of the most implementing security solutions in network environments to detect anomalies. The major challenge of this kind of this system is to maximize the detection and accuracy rates and minimize false positive.**

The well-known genetic algorithm is based on gene reproduction and mutation. Recent research has pointed out that additional information embedded alongside individual chromosomes transmits data into future offspring. This additional transmission of information into child generations outside DNA is known as epigenetics. Additional information is considered as the epigenetic factor that helps us to define randomness crossover and mutation used in classical genetic algorithm.

This paper presents a state of art where we try to explore epigenetic algorithms within the context of Intrusion Detection System. We discuss the methodology used in genetic algorithm and how our approach can perform detection of intrusions for an efficient security.

Keywords— Epigenetic algorithm, genetic algorithm, intrusion detection, security

I. INTRODUCTION

Intrusion Detection Systems (IDS) is considered as an important solution to detect malicious network attacks. There are two main approaches to detect intrusions. The first is to detect known attacks by known signatures. The second is to define the normal behavior of the system by using heuristic approaches to identify abnormal behavior.

However, the challenge of these detection techniques is to have the best performance of detection accuracy [1]. The major performance metrics developed for intrusion detection are Detection rate and False positive rate. Detection rate (DR) is the ratio between number of correctly detected intrusions and the total number of intrusions, its value should be high. False Positive (FP) (also said false alarm) rate is the ratio between numbers of normal connections that are incorrectly classified as intrusions and the total number of normal connections, its value must be minim.

Several heuristic algorithms are solving problems that deterministic and mathematical methods have not succeed to solve.

Among these heuristic algorithms, we find genetic algorithm that is used to optimize multiple problems in several areas such as mechanics, medicine, finance ... and in the IDS.

Actually, the new challenge is to optimize and make hybridizations of several heuristic algorithms to get the best solution. New research has been done on genetic algorithm by adding additional factors. Researches speak about epigenetic that study gene expression without changing the gene sequence. In this context, we talk about additional operators: epimutation and epicrossover added to the classic genetic algorithm that help us to converge towards a better solution and reduce the number of iterations. Epigenetics is a mechanism controlling genes that are activated or non-activated of chromosome structure; in other words, it controls gene activity.

Epigenetic algorithms [2] are used to resolve too many problems in a several domains. For example, it has been applied in the field of optimization and planning of GSM mobile frequencies [3].

In our state of art, we present in the first section the background and definitions of intrusion detection systems, genetic algorithm, and epigenetic algorithm. Then, we will try to explore epigenetic algorithms within the context of Intrusion Detection System and present our approach. Finally, the last section is dedicated to conclude our work.

II. BACKGROUND AND DEFINITIONS

A. Intrusion Detection System solution

Intrusion detection systems are software or hardware systems that automate the process of monitoring the events occurring in a computer system or network, they analyze them for malicious activities or policy violations and produce reports to a management station [4].

This system includes a set of information used to detect intrusions in a kind of knowledge base attack.

IDS use different techniques: Signature based on detection used to detect known attacks. Anomaly based approach involving collection of data related to a legitimate behavior and then apply statistical tests to the observed behavior, which determines whether that behavior is legitimate or not [5].

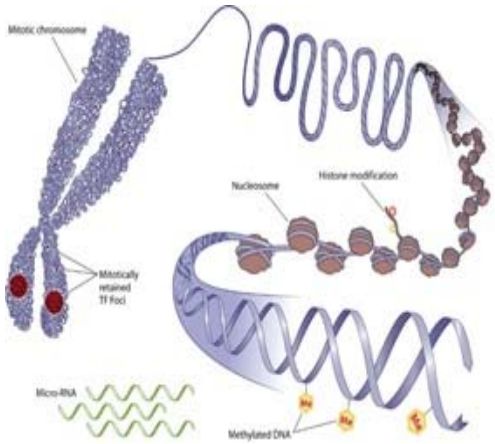
978-1-5090-6227-0/16/$31.00 ©2016 IEEE

There are several computing techniques that can be used to improve detection accuracy of IDS such as Artificial Neural Network (ANN) (Han and Kamber, 2006), Fuzzy logic (Han and Kamber, 2006), Association rule mining, Support Vector Machine (SVM) ((Han and Kamber, 2006), Genetic Algorithm (GA) (Dhanalakshmi and Ramesh Babu, 2008; Li, 2004).

B. Epigenetic Discovery

Epigenetic studies the cellular and physiological phenotypic treating variations caused by external or environmental factors that switch genes on and off and affect how cells can read genes. This field [6] evolved to include any process that alters gene activity without changing the DNA sequence.

Fig. 1. Mechanisms of inheritable epigenetics [7]



As defined by [8, 9] Epigenetic is defined as a term used to specify various processes leading to prolonged modifications in gene expression without making any change on the genetic code, namely DNA base sequence. Recent studies have revealed that epigenetics have a role in the arrangement of gene expressions not only in the growing process but also in adult life.

Many studies has showed that additional information embedded alongside individual chromosomes transmits data into future offspring. This additional transmission of information into child generations outside DNA is known as epigenetics. Additional information that called the epigenetic factor can help us to define the randomness crossover and mutation used in classical genetic mechanism.

The Epigenetic characters are factors that reflect the genetic differences between communities and clarify the manifestations of genes that affect the evolution of people [10]. Through epigenetics, the gene activity is regulated.

An epigenetic modification enables to regulate gene expression according to several conditions. In other words, by changing environmental conditions and with external interventions such as medicine and therapy, it is possible to control gene expression. According to many studies established about epigenetic regulations, it is revealed that environmental factors transfer their impacts to the genome by modifying the gene expression.

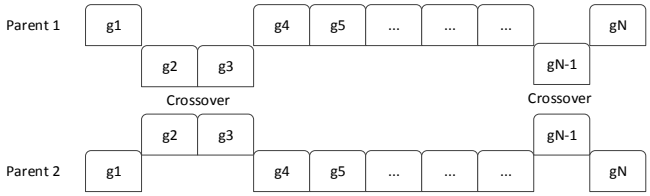
III. GENETIC ALGORITHM APPLIED ON INTRUSION DETECTION

Genetic algorithms [11, 12] is inspired from biology and genetics to iteratively evolve a population of initial individuals to a population of high quality individuals where each individual represents a solution of the problem to be solved and is composed of a fixed number of genes. The number of possible values of each gene is called the cardinality of the gene. Each individual is represented as chromosome that forms population.

Fig. 2. Structure of one chromosome with N genes

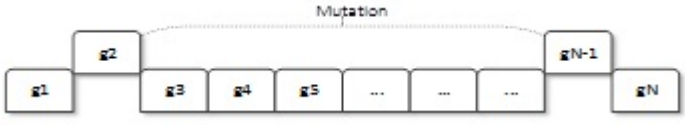
Genetic algorithm begins randomly by generating population of individuals or chromosomes. In every generation, there are three basic operators of genetic algorithm: selection, crossover, and mutation, which are applied to each individual. Crossover Operators is the operator that provides the creation of offsprings with an interchange of gene structures of two chromosomes (parents) which come together. Before the interchange of genes (information) by crossover, the crossover probability should be determined.

Fig. 3. Crossover operator



Mutation Operator means changing of value of a randomly chosen gene of a chromosome. The chromosome's genes are randomly displaced in the same gene series to increase the diversity of genes.

Fig. 4. Mutation operator



Selecting Operator is operation to decide which chromosomes to be passed down to the next generation and which chromosomes to be disappeared are determined ac-

cording to the magnitude of the fitness value. This evaluation is done thought a Fitness function.

In intrusion detection case, the Fitness function used is [13]:

$$\Phi = \alpha/A - \beta/B$$

α: the number of correctly detected attacks.

A: the total number of attacks in the training dataset**.**

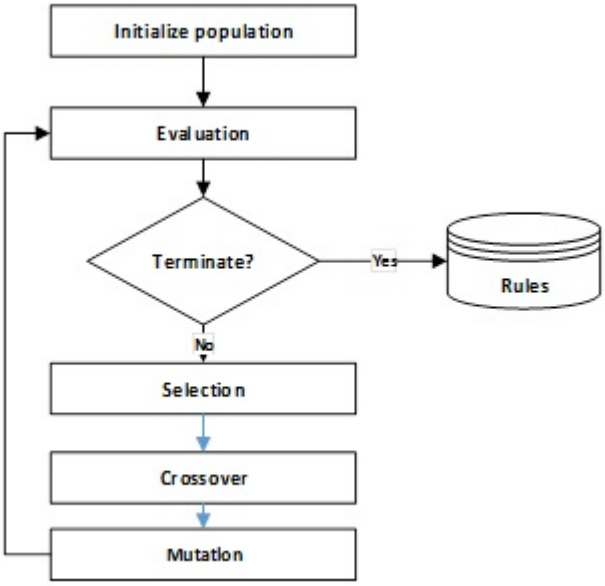
β: the number of normal connections incorrectly characterized as attacks, i.e. false-positives.

B: the total number of normal connections in the training dataset.

Scale of fitness values is [-1, 1]. The high detection rate and low rate of false-positives result is a high fitness value. The low detection rate and high rate of false-positives result is a low fitness value.

The goal of the algorithm is to have the best generation that will define the best solution. Figure 5 shows the genetic algorithm [14].

Fig. 5. Algorithm genetic design



For IDS, genetic algorithm is used to evolve new rules. The role of a rule is to specify whether the source traffic is normal or abnormal. The general syntax of rules [18]:

if {condition} then {act}

Condition is a set of data needs to be checked and while act refers to the decision if this set is attack or normal source. A condition can check for port numbers of network protocols, protocols used, duration of connection, IP address of source and destination etc.

Example of data set is KDD-NSL [15] that is used by researchers to compare different detection intrusion methods. It consists of 4900000 single connection instances. Each connection instance contains 41 features (duration, protocol_type, service, hot flag …) [15]. In genetic algorithm, the condition is the connection instance, which is the individual or chromosome, and each attribute is considered as a gene code (numerical, binary or alphanumerical).

IV. Epigenetic Algorithm

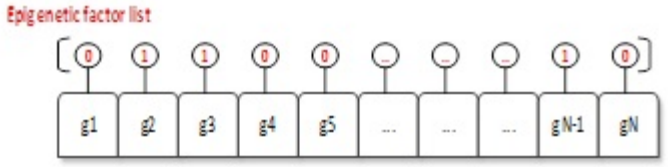
Genetic algorithms are known by mutation and crossover operators that are done randomly by specifying a probability. The new concept of epigenetic algorithm is to specify the genes that are involved in these operators without probability. Inheritance, crossover and mutation operations exist in Epigenetic algorithm structure, they are called as epigenetic inheritance, epicrossover and epimutation [16–17]. Environmental factors, medicine, and other external factors provide the control of the epigenetic factors in epigenetics design.

In epigenetic design, an epigenetic factor list is constituted for each gene receiving a penalty point or not receiving penality and define if this gene is active or not active (we assigned 1 for active gene and 0 for inactive gene). Epigenetic factors are kept as a list throughout all of the population and each chromosome. This list will help us to select genes that will participate in mutation and crossover operation of classical genetic algorithm and in this way; we can reduce the number of iterations of the algorithm.

In addition to the crossover and mutation operators in classical genetic algorithm design, epicrossover and epimutation operators are used in the epigenetic algorithm design proposed in this study.

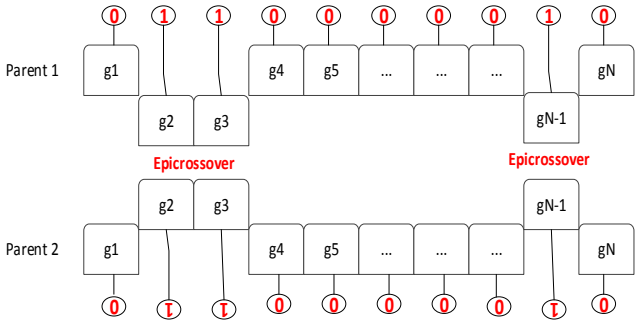
We define at first the epigenetic factor list as showed in figure 6:

Fig. 6. Epigenetic factor list



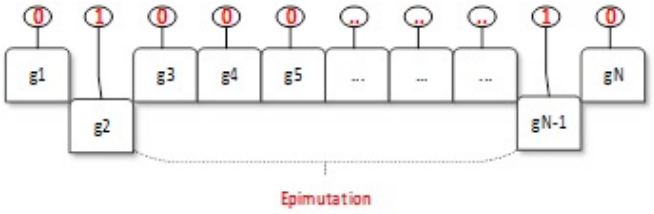
Epicrossover is the new crossover operator that applies for genes having the value ‘1’. This case is represented in figure 7:

Fig. 7. Epicrossover operator



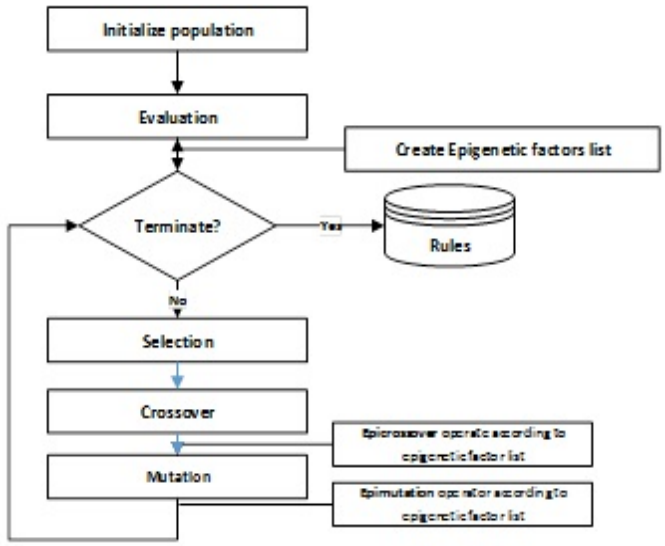
Epimutation is the new mutation operator that applies for genes having the value '1'. This case is represented in figure 8.

Fig. 8. Epimutation operator



The epigenetic algorithm is represented as follow:

Fig. 9. Epigenetic Algorithm design



V. Proposed Approch : Epigenetic Algorithm applied on IDS

In our study for detection intrusion problem, we try to define the epigenetic factor list.

In intrusion detection field, each gene represents the value attributed listed in the table I. These genes constitute the connection or the condition (chromosome) [18].

Connection: {duration = "0:0:1" and protocol = "finger" and source_port = 18989 and destination_port = 79 and source_ip = "99.19.99.19" and destination_ip = "192.168.254.10"….}

TABLE I. RANGE VALUE OF SOME ATTRIBUTE (GENE)

At-tribue name	Range of values	Attribue name	Range of values
Source bytes	0.0.0.0~255.255.255.255	Protocol	1~9
Destination bytes	0.0.0.0~255.255.255.255	Number of Bytes Sent by Originator	0~99999 99999
Source Port Number	0~65535	Number of Bytes sent by Responder	0~99999 99999
Destination Port Number	0~65535	State	1~20
Duration	0~99999999	Service	DNS, FTP, SNMP…

In Dataset KDD_NSL, there are 41 features for each connection that are detailed in Table II [19].

TABLE II. RANGE VALUE OF SOME ATTRIBUTE (GENE)

	Attribute name		Attribute name
1	duration	**22**	is_guest_login
2	protocol_type	**23**	count
3	service	**24**	srv_count
4	flag	**25**	serror_rate
5	src_bytes	**26**	srv_serror_rate
6	dst_bytes	**27**	rerror_rate
7	land	**28**	srv_rerror_rate
8	wrong_fragment	**29**	same_srv_rate
9	urgent	**30**	diff_srv_rate
10	hot	**31**	srv_diff_host_rate
11	num_failed_logins	**32**	dst_host_count
12	logged_in	**33**	dst_host_srv_count
13	num_compromised	**34**	dst_host_same_srv_rate
14	root_shell	**35**	dst_host_diff_srv_rate
15	su_attempted	**36**	dst_host_same_src_port_rate
16	num_root	**37**	dst_host_srv_diff_host_rate
17	num_file_creations	**38**	dst_host_serror_rate
18	num_shells	**39**	dst_host_srv_serr

	Attribute name		Attribute name
			or_rate
19	num_access_files	**40**	dst_host_rerror_rate
20	num_outbound_cmds	**41**	dst_host_srv_rerror_rate
21	is_host_login		

According KDD_NSL, the attacks are categorized into four categories [16]:

- **DoS** (Denial Of Service) attack aim to make a service or a resource unavailable.
- **U2R** (User to Root) attack witch a simple user tries to exploit a vulnerability in order to obtain super user or administrator privileges
- **R2L** (Remote to Local): the attacker attempts to gain access (account) locally on a machine accessible via the network.
- **PROBE** represents any attempt to collect information about the network, the users or the security policy in order to outsmart it.

For each category, we estimate values for attributes in the attack and create a list by assigning the value '1' (active gene) when participating in an attack and value ' 0 '(inactive gene) when not participating in an attack. For example, we can classify the attributes of 'Source IP address' as an active gene '1' if it is in a geographically suspect address pool. We can classify the 'duration' attributes as active gene '1' if the value is above a threshold defined as the attack 'Slow and Low'...etc.

TABLE III. CREATING EPIGENTIC FACTOR LIST IN IDS FIELD

Gene (attribue)	Threshold	Epigene (0,1)
Source bytes	Range of @IP ∈ {geographically country, gov, bank...}	1= true Else 0
Destination bytes	Range of @IP ∈ {geographically country, gov, bank...}	1= true Else 0
Duration	≥ threshold	1= true Else 0
Destination Port Number	Suspect port number (23 (DNS) for DDoS attack…)	1= true Else 0

Also, we can select attributes that participate to each class of attack using methods of selection and correlation attributes. For example, Fisher Linear Discriminant Analysis Method (FLDA) used in research Giffy & Ravichandran [21] allowed to select the Attributes listed in the table IV below :

TABLE IV. SELECTED ATTRIBUTES [21]

No	Attack Category	Selected Attributes
1	DoS	failed logins logged in Count Same srv rate srv diff host rate dst host srv diff host rate
2	Probe	logged in serror rate rerror rate same srv rate dst host srv count dst host srv diff host rate dst host serror rate
3	U2R	Urgent Root shell #shells Is hot login dst host srv count dst host serror rate
4	R2L	protocol type service flag source bytes failed login logged in #shells #access files is hot login is guest login srv count rerror rate srv rerror rate srv diff host rate dst host serror rate dst host rerror rate dst host srv rerror rate

We assign value (1) to these attributes (genes) that participate in the detection of an attack and the values (0) to others genes that do not participate in the detection of an attack.

After defining the list of active and inactive genes, we create a list of epigenetic factors to apply epigenetics algorithm.

EpiGenetic Algorithm

{

Create **Initial Population (initial connections)**

do

{

Evaluate the chromosomes according to Fitness function

Set to epigenetic factors to genes (**epigenetic factors list (according one of selection method for example)**)

Apply **Selection**

Apply **Crossover operator**

Apply **Epicrossover** operate according to epigenetic factor list

Apply **Mutation operator**

Apply **Epimutation** operator according to epigenetic factor list

Create **Next population** from previous population

} **while** (!(end of **Generation**) **or** !(stop to get better))

Take **the best** chromosome to solution

}

This algorithm will reduce the number of searches and iterations to converge to an optimal solution by applying the mutation and crossover operations without random.

VI. CONCLUSION AND FUTURE WORKS

Epigenetics is a recent and an important innovation in the field of biology. It helps to prevent more precisely the curable and not curable diseases based on environmental factors that do not fit in the sequence gene.

From this biological inspiration, we try to apply this concept to the intrusion detection field and look for epigenetic factors corresponding to converge to an effective solution an efficient security.

In our future works, we will look to experiment our approach and compare the results with existing genetic algorithm to demonstrate the reduction of total iterations in the aim to obtain the optimal solution in a shorter time.

REFERENCE

[1] Robin Sommer, Vern Paxson : "Outside the Closed World: On Using Machine Learning For Network Intrusion Detection"

[2] Computational Methods in Epigenetics, Vanessa Aguiar-Pulido Jose M. Eirin-Lopez, Javier Pereira, Giri Narasimhan, Victoria Suarez-Ulloa, Chapter 6

[3] EpiGenetic Algorithm for Optimization: Application to Mobile Network Frequency Planning. King Fahd University of Petroleum & Minerals 2015 (DOI 10.1007/s13369-015-1869-5)

[4] Miller, Brad. L. and Michael J. Shaw. 1996. "Genetic Algorithms with Dynamic Niche Sharing for Multimodal Function Optimization." In Proceedings of IEEE International Conf. on Evolutionary Computation, pp. 786-791. Nagoya University, Japan

[5] C. Modi, D. Patel a, B. Borisaniya, H. Patel, A. Patel, M. Rajarajan "A survey of intrusion detection techniques in Cloud". Elsevier Journal ofNetworkandComputerApplications36(2013)42–57

[6] Maynard, S.J.: Models of a dual inheritance system. J. Theor.Biol. 143, 41–53 (1990)

[7] A. Zarrabi, A. Zarrabi "Internet Intrusion Detection System Service in a Cloud". IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 2, September 2012

[8] ASM (American Society for Microbiology)

[9] Delcuve, G.P.; Rastegar, M.; Davie, J.R.: Epigenetic control. J.Cell. Physiol. 219, 243–250 (2009).

[10] AN IMPLEMENTATION OF INTRUSION DETECTION SYSTEM USING GENETIC ALGORITHM, International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.2, March 2012

[11] Corolineberry, A.C.; Berry, R.J.: Epigenetic variation in the human cranium. J. Anat. 101, 361–379 (1967)

[12] S. Selvakani and R.S. Rajesh, "Genetic Algorithm for Framing Rules for Intrusion Detection" IJCSNS International Journal of Computer Science and Network Security, Vol. 7 No. 11, November 2007– 8 Cubas, P.; Vincent, C.; Coen, E.: An epigenetic mutation responsible for natural variation in floral symmetry. Nature 401, 157– 161 (1999)

[13] Kirkpatrick, B.: Computer algorithm uses epigenetics to identify aging genes http://www.whatisepigenetics.com/computeralgorithm-uses-epigenetics-to-identify-aging-genes. (2014)

[14] Improved Genetic Algorithm for Intrusion Detection System, 2014 Sixth International Conference on Computational Intelligence and Communication Networks .

[15] Intrusion Detection System Using Genetic Algorithm, Science and Information Conference 2014 August 27-29, 2014 | London, UK.

[16] A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection, International Journal of Engineering Research & Technology (IJERT), International Journal of Engineering Research & Technology (IJERT)

[17] A. Zarrabi, A. Zarrabi "Internet Intrusion Detection System Service in a Cloud". IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 2, September 2012

[18] McHugh, John, 2001. "Intrusion and Intrusion Detection." Technical Report. CERT Coordination Center, Software Engineering Institute, Carnegie Mellon University

[19] Efficient Classifier for R2L and U2R Attacks, International Journal of Computer Applications (0975 – 8887) Volume 45– No.21, May 2012. P. Gifty Jeya, M. Ravichandran, C. S. Ravichandran.

[20] A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms. International Journal of Advanced Research in Computer and Communication Engineering (Vol. 4, Issue 6, June 2015) L.Dhanabal , Dr. S.P. Shantharajah.

[21] Epigenetics: The Science of Change : Articles from Environmental Health Perspectives are provided here courtesy of National Institute of Environmental Health Science.

[22] V. Moraveji Hashemi, Z. Muda and W. Yassin, 2013. Improving Intrusion Detection Using Genetic Algorithm. Information Technology Journal, 12: 2167-2173.