Open access Journal International Journal of Emerging Trends in Science and Technology

Wrapper Approach for IDS Using Fuzzy Genetic Algorithm

Authors
Pooja U. Chavan[1], Priyadarshani K. Chavan[2], Rutuja S. Choudhari [3]
Vidya A. Gaikwad[4]
[1]Bharati Vidyapeeth's College of Engineering for Women, Katraj,Pune-43,
Email: *chavanpooja093@gmail.com*
[2]Bharati Vidyapeeth's College of Engineering for Women, Katraj,Pune-43,
Email: *chavanpriyadarshani11@gmail.com*
[3]Bharati Vidyapeeth's college of Engineering for Women,Katraj,Pune-43,
Email: *rutuja.star100@gmail.com*
[4]Bharati Vidyapeeth's College of Engineering for Women,Katraj,Pune-43,
Email: *gaikwadvidya623@gmail.com*

Abstract
The intrusion detection systems (IDS) are becoming indispensable for effective protection against attacks that are constantly changing in magnitude and complexity. This paper proposes a fuzzy genetic algorithm (FGA) for intrusion detection. The FGA system is a fuzzy classifier, whose knowledge base is modeled as a fuzzy rule such as "if-then" and improved by a genetic algorithm. The method is tested on the benchmark KDD'99 intrusion dataset and compared with other existing techniques available in the literature. The results are encouraging and demonstrate the benefits of the proposed approach.
Keywords: *classification,DARPA data set, fuzzy logic,genetic algorithm , intrusion detection*

Introduction

With the development of Internet, intrusion detection systems (IDS) have received remarkable attention.

An intrusion detection system is software or hardware or both of them designed to monitor computer system or network activities for malicious activities or policy violations. Intrusion detection is an important topic in computer security. There are two main intrusion detection models: anomaly detection and misuse detection approaches

The anomaly detection model describes the usual behaviour of a user to detect this user's anomalous or unaccustomed action. Among methods proposed to construct profiles, we mention: the statistical methods where the profile is calculated from variables taken randomly and sampled at regular intervals. These variables can be, for example, the number of connections, the number of erroneous passwords, etc. The expert systems and neural networks are two well-known methods used to calculate a user profile.

The misuse detection model defines some anomalous behaviour to analyze data susceptible to be attacks. The approach often uses known attacks called signatures. Among these methods, we mention: the expert systems,the genetic algorithm and the pattern matching method that provides signatures of attacks. Various algorithms are used to localize these signatures in the audit trail.

Recently, several systems have been built to detect intrusions.Various techniques have been applied extensively for intrusion detection such as agents-based detection intrusion which can provide many advantages for the existing solutions due to the mobility of agents and their cooperative aspects, the Data mining approaches,the clustering techniques

,the naïve Bayesian classifier and the fuzzy evolutionary algorithms .Fuzzy logic .is an intelligent method that has been successfully employed for many IDSs

In this work, we focus on fuzzy genetic algorithms for intrusion detection. The methodology is a combination of the genetic algorithm with the fuzzy logic concepts.

Genetic algorithms provide a natural tool to solve several problems in the field of applied mathematics and science in general. Thus by combining genetic algorithms with fuzzy logic formalism we obtain complete and consistent enough for the acquisition, representation and use of knowledge by computers. We used the concept of fuzzy logic in solving the problem of intrusion detection because fuzzy logic is an effective tool for introducing the concept of membership degree that determines the "strength" in which an object belongs to different classes.

The paper is organized as follows. The second section gives an overview of the DARPA dataset. The third section presents the fuzzy genetic algorithm for intrusion detection. The implementation and some numerical results are given in the fourth section. Finally, the fifth section concludes the work.

AN OVERVIEW OF DARPA INTRUSION DATASET

The oriented intrusion detection dataset used in the experimental study of this work are those of KDD'99 (http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html).

As shown in TABLE I, the KDD99 dataset contains 22 different attack types which could be classified into four main categories namely Denial of Service (DOS), Remote to User (R2L), User to Root (U2R) and Probing.

Table I. Types Of Attacks In Kdd'99 Dataset

Main Attack Classes	**22 Attacks Classes**
Denial of Service(DOS)	*Bac,land,neptune,pod,smurf,teardrop*
User to Root(U2R)	*Bufferoverflow,loadmodule,pearl,rootkit*
Remote to User(R2L)	*ftpwrite, guess password, phf, spy*
Probing	*Ipsweep, nmap,portsweep, satan*

The full DARPA dataset contains 4885950 lines of connections. TABLE II gives the percentages and the number of connections per class.

Table II.The No Of Connections In Each Class

Normal	***DOS***	***U2R***	***R2L***	***Probing***
972781	3883370	52	1126	28621

Table III.The Percentage Calculated For Each Class

Normal	***DOS***	***U2R***	***R2L***	***Probing***
19.91%	*79.48%*	*0.00001 %*	*0.00023 %*	*0.0059 %*

THE PROPOSED APPROACH

The proposed approach consists of two main steps. The first one is the data normalization. The second step is the fuzzy genetic algorithm.

A. Data-preprocessing and Normalization

Each line of the KDD'99 dataset called "connection" includes a set of 41 features and a label which specifies the status of connection as either normal or specific attack type.

The features of a connection include the duration of the connection, the type of the protocol (TCP, UDP, etc), the network service (http, telnet, etc), the number of failed login attempts, and the service and so on. These features had all forms of continuous, discrete, and symbolic, with significantly varying ranges. Among the 41 attributes of the connection, we consider only sixteen significant attributes which are: A8, A9, A10, A11, A13, A16, A17, A18, A19, A23, A24, A32, A33, A1, A5 and A6. These attributes are normalized. The normalization formula given in (1) is applied in order to set

attribute numerical values in the range [0.0, 1.0].

$$X = \frac{X - MIN}{MAX - MIN} \quad (1)$$

Where X: is the numerical attribute value, MIN is the minimum value that the attribute X can get and MAX is the maximum one.

Significant attributes are the important ones that can help in classifying a connection correctly.

After having analyzed the KDD'99 dataset, the MIN and MAX values of each significant attributes which we have selected and considered in the current work are given as follows:

- A8: is the number of ``wrong'' fragments, values in the range [0.3] (MIN = 0 MAX = 3),
- A9: is the number of urgent packets,values in the range of [0,14,],
- A10: is the number of "hot" indicator,values in the range [0.101],
- A11: is the number of failed login attempts, values in the range [0,5];
- A13: is the number of ``compromised'' conditions, values in the range [0.9],
- A16: is the number of "root" accesses,values in the range [0.7468],
- A17: is the number of file creation operations, values in the range [0,100],
- A18: is the number of shell prompts, values in the range [0,5]
- A19: is the number of operations on access control files, values in the range [0.9],
- A23: is the number of connections to the same host as the current connection in the past two seconds, values in the range [0.511],
- A24: is the number of connections to the same service as the current connection in the past two seconds, values in the range [0.511],
- A32: is the number of connection to the same host, values in the range [0,255]
- A33: is the number of connection to the same serves for the host, values in the range [0,255].
- A1: duration is number of seconds of the connection, values in the range [0. 58329].
- A5: is the number of data bytes from source to destination, values in the range [0.1.3 one billion].
- A6: is the number of data bytes from destination to source, values in the range [0. 1.3 one billion].

However, for the numerical attributes A1, A5 and A6, we have observed a big value of MAX hence the need tomodify the normalization formula given in (1). The logarithmicscaling (with base 10) is applied to these features to reduce the range.

We used all the sixteen features as the inputs of our Local-fuzzy classifier which is detailed in the next section.

B. The fuzzy genetic algorithm

The Fuzzy genetic algorithm (FGA) starts from a population of individuals generated randomly. Each individual is an " if-then" fuzzy rule. In order to optimize the set of fuzzy rules already generated in the first stage, a genetic algorithm process which consists of selection, crossover and mutation operators are applied on the individuals.

C. The fuzzy rule encoding

A fuzzy rule "if-then" is encoded as a string. We have used a vector of 16 bits where each bit corresponds to an attribute. Five possible linguistic values may be used for each attribute which are: S: Small, MS: Medium Small, M: Medium, ML: Medium Large and L: Large. Figure 1 draws the Membership functions of the five linguistic values.

For example:

- Let us consider the rule: If X1 is medium, X2 is medium small X3 is large and X4 is small, then Class= Cj with CF = CFj. Where Xi is the connection attribute, Cj is the class obtained after classification and CFj is its degree of confidence. The Corresponding code is : "M, MS, L, S"

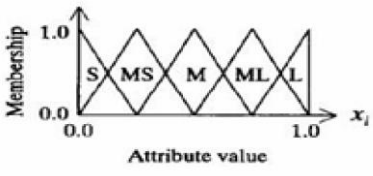

Figure 1.Membership functions of five linguistic values (S: small, MS: medium small, M:

medium, ML: medium large, L: large).

D. The membership function μ (X)

The membership function for each attribute X noted μ(X) is calculated by a projection on the graph of the fuzzy set depicted in Figure 2. Formula (2) shows how we can calculate the μ(X) value.

$$\mu(x) = Max\{0,1-\left(\left|X - X0\right| / b\right)\} \quad (2)$$

where: b is the base of the triangle, b = 0.5.
X0 = {0, 0.25, 0.5, 0.75, 1} corresponding to {S, MS, M, ML, L}. X: is the attribute value after normalization

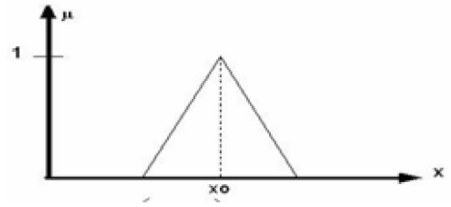

Figure 2. The Fuzzification method

E. An individual representation

The individual representing a fuzzy " if-then" rule is generated randomly. For each attribute Xi, a linguistic value (among the five values of the fuzzy set) is assigned randomly.

F. The evaluation of a fuzzy rule and the fitness function

To evaluate an " if-then" rule Rj and classify a connection Xp with a certain confidence degree, we have used the method introduced in.
To evaluate a fuzzy rule Rj, we give the following steps:

- Calculate the compatibility of connections with the rule Rj: Let us consider the fuzzy if-then rule Rj denoted

"A A ... ."A ", we calculate the compatibility of j1,J2 In each connection Xp of the dataset with the rule Rj by using the Formula (3).

$$\mu Rj(Xp) = \mu Aj1(X1) \times \mu Aj2(X2) \times \ldots \mu Ajn(Xn), P = 1,2,\cdots n \quad (3)$$

where μ () is the membership function. m: is the total number of connections. Xi: are the attributes. Xp is the current connection and n is the number of attributes which equals to 16.

- Calculate the sum of the compatibilities for each class of the five categories: for each class h belonging to the five classes DoS, R2L, U2R, Probing and Normal, we calculate the sum of compatibilities as given in Formula (4).

$$\beta CLASSh(Ri) = \sum_{Xp \in CLASSh} \mu Rj(Xp) \; h = 1,2,\cdots c \; c = 5 \quad (4)$$

- After having calculated the sum of compatibilities of a rule Rj for each class h, we selected the class having the maximum value (as given in Formula (5)). This class Cj is considered the suitable class for the rule Rj. If two classes had the same maximum value then the class is not specified (Cj = null) and CFj = 0.

$$\beta CLASScj(Rj) = \max\{\beta CLASS1(Rj) \ldots \beta CLASSc(Rj)\} \quad (5)$$

The confidence degree CFj of the class Cj for the rule Rj is computed by the Formula (6).

$$CFj = \frac{\beta CLASScj(Rj) - \overline{\beta}}{\sum_{h=1}^{C} \beta CLASSh(Rj)}$$

$$\overline{\beta} = \frac{\sum_{h \neq cj} \beta CLASSh(Rj)}{c-1} \quad (6)$$

The formula (7) shows how the fitness of a fuzzy rule is obtained. The PPF represents the Positive Power Rule. The
fitness value of a rule is the sum of the PPF for all considered classes.

$$fitness(Rj) = \sum_{p \in CLASScj} PPF_{\gamma}^{Rj}$$

$$PPF^{Rj} = \begin{cases} 1 \; if \; \mu Rj(Xp) \succ 0 \\ 0 \; otherwise \end{cases} \quad (7)$$

G. The genetic algorithm operators

The genetic algorithm we used performs for each generation:

- A random one-point crossover on two randomly selected individuals.
- A random mutation of all genes of an individual randomly selected. The mutation operator has two functions:

- A regulation of the population explosion caused by the crossover operator.
- The enrichment of the population by introducing new genes.

A selection of individuals having a fitness value >0. So all individuals having a fitness value equals to zero are discarded and eliminated from the population. We consider only individuals with a fitness value superior to zero.

H. The fuzzy genetic algorithm organigram

The different steps of the proposed fuzzy genetic algorithm for intrusion detection are depicted on Figure 3.

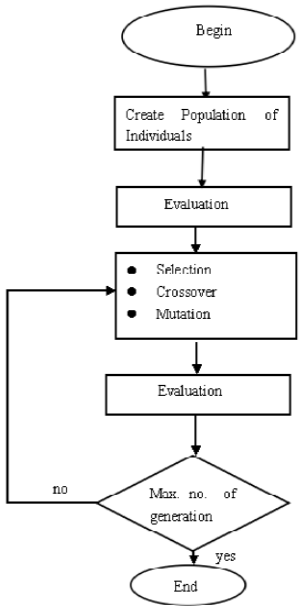

Figure 3. A fuzzy genetic algorithm for intrusion detection (FGA).

EXPERIMENTAL RESULTS

The implementation was done on MATLAB. First, we have created five matrices: the matrix containing the U2R-events, the matrix containing R2L-events, the matrix containing the Probing events, the matrix containing the DOS events and the matrix containing the normal connections. Then, the normalization phase is launched where the various attributes of connections of all matrices are normalized We have obtained five normalized matrices U2R, R2L, Probing, Normal and DOS. The next step is the generation of fuzzy rules. To do this, we used the "rand" functio n (random number to generate random numbers that must be among the five values (1, 2, 3, 4. 5) which correspond to (Small, Medium Small, Medium, Medium Large and Large).

We have applied the FGA on the five matrices Rand representing fuzzy rules. All experiments were performed on a laptop CPU Core 2 Duo 2.0 Ghz (x 2) with 3 Go of Ram. To evaluate the performance of the approach, we used the following measures:

[10] True Positives (**TP**): is the number of normal connections classified by the genetic approach as normal.

[11] False Positives (**FP):** is the number of normal connections classified as attacks by the genetic approach.

[12] True Negatives (**TN)**: is the number of attack connections classified as attacks by the genetic approach.

[13] False Negatives (**FN)**: is the number of attack connections classified as normal by the genetic approach.

[14] Specificity: It describes the ability to identify negative results (test the reliability of the given method).

$$specificity = \frac{TN}{TN + FP}$$

[15] Sensitivity: It describes the ability to identify positive results

$$Sensitivity = \frac{TP}{TP + FN}$$

TABLE IV to TABLE VIII give the performance of the proposed FGA method applied to the five classes. According

To the results obtained, we have observed that the FGA successeds in finding good results for the five classes DOS,R2L,U2R, Probing and Normal, and false alarms are minimal.

The success rates are as follows:

[16]**99,99%** for DOS class,
[17]**92,5%** for Normal class,
[18]**92,5%** for R2L class,
[19]**92,5%** for U2R class,
[20]**92,5%** for Probing class.

Table IV.The Performance Measure Of Normal Class

TP	***FP***	***TN***	***FN***
375	*2*	*1131*	*2*
specificity	*FP rate*	*sensitivity*	*FN rate*
0.9982	*0.0018*	*0.9947*	*0.0053*

Table V. The Performance Measure Of U2r Class

TP	***FP***	***TN***	***FN***
58	*2*	*1171*	*0*
specificity	*FP rate*	*sensitivity*	*FN rate*
0.9983	*0.0017*	*1*	*0*

Table VI.The Performance Measure Of R2l Class

TP	***FP***	***TN***	***FN***
46	*2*	*1171*	*0*
specificity	*FP rate*	*sensitivity*	*FN rate*
0.9983	*0.0017*	*1*	*0*

Table VII. The Performance Measure Of Dos Class

TP	***FP***	***TN***	***FN***
557	*2*	*1139*	*2*
specificity	*FP rate*	*sensitivity*	*FN rate*
1	*0*	*0.9964*	*0.0036*

Table VIII.The Performance Measure Of Probing Class

TP	**FP**	**TN**	**FN**
97	2	1171	0
specificity	FP rate	sensitivity	FN rate
0.9983	0.0017	1	0

A. Comparative Study

In order to situate our contribution, we compare our results with some well-know methods for intrusion detection such as: FLS [5], Hybrid EFS [2], C4.5 [18], 5-NN [3], EFRID (Evolving Fuzzy Rules for Intrusion. Detection) proposed in [10], NB [13] and Naive Bayesian classifier [4]. TABLE IX presents the results obtained for the five classes.

Table IX. Comparision Of Some Algorithms

	CLASS	**CLASS**	**CLASS**	**CLASS**	**CLASS**
Algorithm	***Normal %***	***U2R %***	***R2L %***	***DOS %***	***Probing %***
FGA	*92.5*	*92.5*	*92.5*	*99.99*	*92.5*
FLS	*10*	*95*	*85*	*80*	*80*
Hybrid EFS	*98.5*	*96.3*	*89*	*98.5*	*82.5*
C4.5	*95.9*	*21.1*	*30.2*	*97.1*	*96.3*
5-NN	*96.3*	*25.4*	*3.8*	*96.7*	*87.5*
EFRID	*92.78*	*13*	*7.45*	*98.91*	*50.35*
NB	*94.2*	*25*	*5.4*	*79.4*	*90.4*
Naive bayesian	*97.68*	*11.84*	*8.66*	*96.65*	*88.33*

From TABLE IX, it can be seen that interesting result are obtained. For all the five classes U2R, R2L, DOS, Probing and Normal, the FGA finds good results compared to the other methods.

To further illustrate the results of TABLE IX, we give the comparative curves in Figure 4 to show the effectiveness of FGA in reaching good quality solutions compared to FLS, hybrid EFS, C4.5, 5-NN, EFRID, NB and Naïve Bayesian for intrusion detection.

Acknowledgment

F.A. Authors thank our management for the support and encouragement for carrying out this work.

References

1. Shingo Mabu, Member, IEEE, Ci Chen, Nannan Lu, Kaoru Shimada,and Kotaro Hirasawa, Member, IEEE" An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming" TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 41, NO. 1, JANUARY 2011.
2. P.Jongsuebsuk+,N.Wattanapongsakorn+, C. Charnsripinyo* +Department of Computer EngineeringKing Mongkut's University of Technology Thonburi, Bangkok, Thailand." Real-Time Intrusion Detection with Fuzzy Genetic Algorithm" **978-1-4799-0545-4/13/$31.00 ©2013 IEEE**
3. Mostaque Md. Morshedur Hassan LCB College, Maligaon, Guwahati, Assam, India." CURRENT STUDIES ON INTRUSION DETECTION SYSTEM, GENETIC ALGORITHM AND FUZZY LOGIC"International Journal of Distributed and Parallel Systems (IJDPS), Vol.4, No.2, 2013
4. Ravi Prakash Reddy I, Ph.D Professor & Head in Department of IT GNITS ShaikPet Hyderabad. "A Literature Survey and Comprehensive Study of Intrusion Detection"International Journal of Computer Applications (0975 – 8887) Volume 81 – No 16, November 2013.
5. Prof. P. B. Kumbharkar Research Scholar, JJT University, Rajasthan, India," Intrusion Detection System with Supervised Learning Algorithms", Volume 4, Issue 4, April 2014 ISSN: 2277 128X.

Authors Profile

Miss. Pooja Udhavrao Chavan, Computer Engineering, Savitribai Phule Pune University, Email id:*chavanpooja093@gmail.com*

Miss. Priyadarshani Kishanrao Chavan, Computer Engineering, Savitribai Phule Pune University, Email id:*chavanpriyadarshani11@gmail.com*

Rutuja Sundarrao Chodhari, Computer Engineering, Savitribai Phule Pune University, Email id: *rutuja.star100@gmail.com*

Miss. Vidya Arjun Gaikwad Computer Engineering, Savitribai Phule Pune University, Email id: *gaikwadvidya623@gmail.com*