Volume 4, Issue 12, December 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Modified Intrusion Detection System using Fuzzy Genetic Algorithm

Yogita Danane*, Thaksen Parvat
Department of Computer Science, Sinhgad Institute of Technology,
Lonavala, S. P. P. U. India

Abstract— computing environment is continually growing and changing with new technology and the Internet. In addition, vulnerabilities in this environment are also steadily increasing. So Intrusion Detection Systems (IDS) have turn out to be an important part in provisions of computer and network security. This paper presents a fuzzy-genetic approach to detecting network intrusion. To implement and measure the performance of the system the KDD99 benchmark dataset is used. The KDD99 dataset is a benchmark dataset that is used in various. Genetic algorithm includes a development and collection that uses a chromosome-like data structure and develop the chromosomes using selection, crossover and mutation operators. Fuzzy rule is a machine learning algorithm that can sort network attack data. The results of the proposed system are measured in terms of accuracy, execution time and memory allocation. Results are compared with the existing system which uses sequential algorithm, genetic algorithm or fuzzy algorithm for intrusion detection.

Keywords— Fuzzy algorithm, genetic algorithm, fuzzy genetic algorithm, intrusion detection system, KDD Cup 1999 dataset

I. INTRODUCTION

The speedy increase in the use of computer and computer network in today's civilization seeks extremely secured and trusted communication. There are a range of approaches being utilized in intrusion detections, but any of the systems is not entirely perfect. Intrusion is referred as any set of events that try to negotiate the integrity, confidentiality, or availability of a computer resource. The process of finding out the irregular activities on the system or the network is known as intrusion detection. There are two methods of detection signature-based and anomaly-based. The signature-based method tries to match machine action to stored signatures of known anomalies or attacks. Anomaly Detection as the approach detects anomalies by first learning the characteristics of normal activity.

Misuse Detection: known intrusions are detected by observing the computer system activities some attribute pattern of such intrusions. This goes toward some collected information about the system activities under standard circumstances and under some known intrusions to decide the existing state of the system. In this type, the intrusion detection problem is a categorization problem.

Anomaly Detection: recognizable and unknown intrusions are detected by analysing changes in the standard model of use or performance of the computer system. This method does not use the information about the system performance when an intrusion is in steps forward. There are two types of IDS as: Host-based and Network Based IDS.

Network Based IDS (NIDS): It tests network traffic to identify threats that produce irregular traffic flows, like Denial of Service (DoS) attacks, scanning, and certain structures of malware.

Host-Based IDS (HIDS): It observes the uniqueness of a solo host and the actions occurring in that host for suspicious action.

In this paper, a fuzzy-genetic approach to detecting network intrusion is proposed. To implement and calculate the performance of the system the KDD99 benchmark dataset and network dataset are used. The fuzzy logic is encoded using six attributes.

The rest of this paper is offered as follows. In section II, Block diagram and algorithms are discussed. Section III presents Network datasets. Then Section IV shows result and analysis. Finally in section V conclusion of the experiment is given.

II. LITERATURE SURVEY

A. Network Intrusion Detection System

Intrusion Detection System (IDS) has become a major study area in Computer-based security. It is a well-known skill for enlightening. It is used to protect data consistency and system accessibility throughout an intrusion. When a person tries to access an information structure in the system or does any illegal action, the action is known as an intrusion that can be divided into two ways, exterior and interior. The exterior is those people who do not have authority to access the system information and still they try to obtain illegitimately with the help of different saturation techniques. While interior are those who have legal access authorization, but desire to carry out illegitimate activities. Methods for the

intrusion may include miss configurations of the machine and, password cracking, sniffing unsecured traffic, or utilizing the particular protocols design flaw.

B. Genetic Algorithm

Genetic Algorithm (GA) is an efficient investigating process used in computing to locate exact or probable solutions to optimization and search problems. GAs is categorized as global search heuristics. In GA heuristic search is based on the concept of biological evolution. In Genetic algorithms iterative mathematical modelling method is used to find the optimal combinatorial state provided a set of parameters of interest. The population evolution process is simulated using genetic programming. The inhabitants of fixed length are evolved with a Genetic Algorithm by applying crossover and mutation operators with a fitness function that concludes how individuals are to reproduce. A testing solution is developed. Each of which is determined to get fitness. By selecting the better of them, a new generation is created. The procedure is continued through numerous generations with the endeavour that the population should evolve to contain a solution that is acceptable.

Li [8] represents a technique using GA to detect abnormal network intrusion. This approach includes getting classification rules for quantitative and distinct features of network data. The implementation of rule generation for Intrusion Detection System is given, but results of experiments do not exist.

Bridges [1]: This method combines both Genetic Algorithm and fuzzy data mining techniques to detect the misuses and network anomalies. The most features are not predicted properly in a variety of existing Genetic Algorithm based Intrusion Detection Systems. This method uses GA to know the optimal parameters of the fuzzy functions to select the features of the relevant network.

Lu [9]: In this way classification rules are generated by Genetic Programming. Detection or Classification of intrusions in a network using the fitness function is fine-tuned by this method. The time necessary to train the system with vast data creates Genetic Programming implementation difficult.

Crosbie [3]: Different agent techniques and Genetic Programming can be used to detecting network intrusions. The set of agents that conclude the network behaviours can be found out by an agent who monitors one parameter of the network check data and Genetic Programming. Many small independent agents can be used in this method that is an advantage and the communication among the agents is a drawback.

This system finds out the attacks using rule set by executing GA, then develop rules only for R2l and DoS type of attacks. From these two attacks, one from each is selected. The collective performance of the system is less than 60%.

III. ALGORITHM

A. Fuzzy Genetic Algorithm

In this paper, a fuzzy-genetic algorithm is the use described in P. Jongsubsuk [6]. There are two phases in the training phase rules will be improved the by using evolutionary concept from a genetic algorithm then in testing phase these rules will be used to classify data. The following diagrams gives the steps in which fuzzy genetic algorithm will execute.

B. Genetic Algorithm

A GA is a probabilistic search algorithm. It iteratively transforms a set called as population of mathematical objects that is fixed-length binary character strings, each with an attached fitness value into a new population of offspring objects. It uses the Darwinian Principle of natural selection and using procedures that are arranged after apparently taking place genetic operations, such as inheritance, crossover, selection and mutation. These algorithms predetermine a possible solution to the exact problem on a simple chromosome like data structure and relate recombination operators to these structures so as to protect critical information. Genetic algorithms are frequently referred as function optimizers even though the variety of problems to which genetic algorithms have been applied is pretty extensive. Genetic Algorithms are good at making huge search spaces and navigating them, in search of optimal combinations of things, solutions.

1) GA_Rule Generation: The algorithm for GA rule generation is as follow

Input:

An input to the Genetic Algorithm is an encoded binary string of length n, where n is the number of features being accepted, population size(), number of generations, crossover probability (Pc), mutation probability (Pm).

Output: The output is a rule set generation for IDS.

- Initialize the population arbitrarily having the size of each chromosome 41.
- Initialize N. Where N is the total number of records in the training set,
- Pc=0.8
- Pm=0.088.
- for each chromosome in the new population
- Calculate $Fitness = \frac{Fx}{sum(Fx)}$
- Select 50% best fit chromosome and after that eliminate worse fit chromosome.
- Apply Crossover to best-selected chromosome.
- Apply Mutation for each chromosome to generate a new population. Go to step no3.
- Stop

2) GA parameters: GA has some common fundamentals, and parameters defined as:

- GA Operators- The different GA parameter are selection, mutation and crossover. These are the most winning parts of the algorithm because they are contributing in the generation of each population.
- Selection phase-In selection phase population individuals with superior fitness are selected, or else it is damaged.
- Crossover- It is a method in each pair of each randomly participates in exchanging their parent's genes with each other, awaiting an entire new population to be generated.
- Mutation- It flips a number of the bits in an individual, and because all bits could be filled, there is a low probability of detecting the alter.
- Fitness Function- It is defined as a function that scales the value of individual relative to the rest of the population. It generates the best likely solutions from the quantity of candidates in the population.

In pre-processing phase, the KDD99 dataset is processed with Weka tool. It is used to remove the redundant data from existing dataset that result in tested dataset. The elimination of redundant data or records from dataset. It improves the detection rate of the preferred result and improves the performance of the system.
The Genetic Algorithm is applied in the detection phase on selected features data set and assigns fitness for each rule with the next fitness function.

$$Fitness = \frac{Fx}{sum(Fx)} \quad (1)$$

Where, sum (Fx) is the total fitness of all individuals Chromosomes, and Fx is the fitness of individual X.

As we all know, all living organisms consist of cells. Each cell includes the same set of chromosomes. Chromosomes are strings of DNA and are served as a replica for the entire organism. A chromosome is made of genes; each gene encodes a protein. Essentially, each gene encodes an attribute. Probable settings for an attribute (e.g. lack, brown) are known as alleles. Each gene possesses its position in the chromosome. This position is known as the locus. A whole set of genetic material that is all chromosomes are called as genome. An exacting set of genes in the genome is referred as the genotype. Every chromosome is an encoding of a solution to the problem. Encoding typically refers to a data structure on behalf of candidate solutions. A population of such chromosomes is operated on by a Genetic Algorithm. Chromosomes possibly will be:

Bit Strings: (0101 .... 1100)
Real Numbers: (43.2 -33.1 .... 0.0 89.2)
Lists of Rules: (R1 R2 .... R22 R23)
Program Elements: (genetic programming)
Permutations of Element: (E11 E3 E7 .... E1 E15)

C. Fuzzy Algorithm

In the Proposed research work data is classified based on different conditions.
Suppose we have two results A1 and A2

IF A1 is medium AND A2 is small
THEN class is 1
IF A1 is medium AND A2 is large
THEN class is 2
IF A1 is large AND A2 is small
THEN class is 2
IF A1 is small AND A2 is large
THEN class is 3

IV. DATASETS

A. KDD99 Dataset

KDD dataset was obtained from the 1998 DARPA Intrusion Detection assessment plan held by MIT Lincoln Labs. The dataset was formed in a replicated in a military network background in which U.S. Air Force LAN which was subjected to replicate attacks.
The KDD99 dataset used is a 20% file which contains Normal, Denial of Service (DoS), Probe connection instances. In this paper, only six attributes are considered from the dataset shown in Table I.

TABLE I SELECTED ATTRIBUTES FROM DATASET

#	**Attribute name**
1	Duration
2	protocol_type
3	service
4	Flag
5	src_byte
6	dst_byte
42	Class

B. Online Dataset

Online dataset is obtained by capturing online network data. A packet sniffer is used to filter the packets on the network. After that, the same procedure used in KDD99 dataset to detect anomalies is followed for online dataset to detect intrusion.

V. RESULTS

KDDCUP99 is developed from DARPA data from MIT Lincoln Laboratory is used to assess Intrusion Detection Systems. In this study, the KDDCUP99 training and testing datasets are used. The results of the proposed system are measured in terms of accuracy, execution time and memory allocation. Results are compared with the existing systems which use sequential algorithm, genetic algorithm or fuzzy algorithm for intrusion detection.

TABLE IIIII SELECTED RESULT OF TRAINING AND TESTING PHASE OF THE PROPOSED SYSTEM FOR VARIOUS ATTACKS TYPES

Attack Type	Trained Data	Test Data
Normal	972781	60593
Dos	3883370	223298
Probe	38786	2793
R2L	12616	6075
U2R	46	39
Total	4907599	292293

Every record of the datasets contains 41 network features and one manually assigned record type. Nine network features were used in the Genetic Algorithm, which are source IP address, destination IP address, connection duration, protocol, source port, destination port, and dst_hst_service_count. Etc. The record type shows whether a record is an abnormal network connection or normal network connection.

Figure I show a comparison between the proposed system and existing system in terms of accuracy of the detection rate of various attacks.

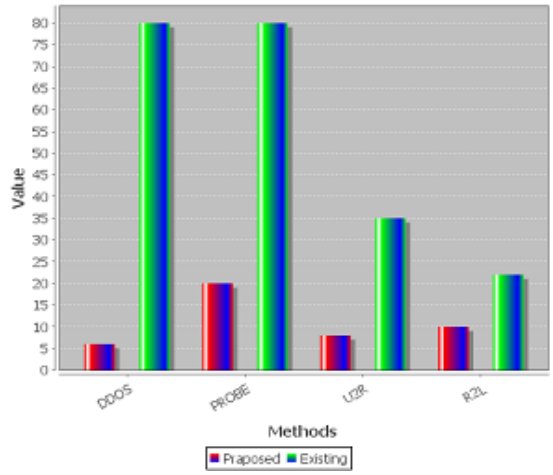

Fig. 1 Comparison between the proposed system and existing systems in terms of accuracy.

Figure.2 shows a comparison between the proposed system and existing systems in terms of time required for training and testing of the dataset on the system.

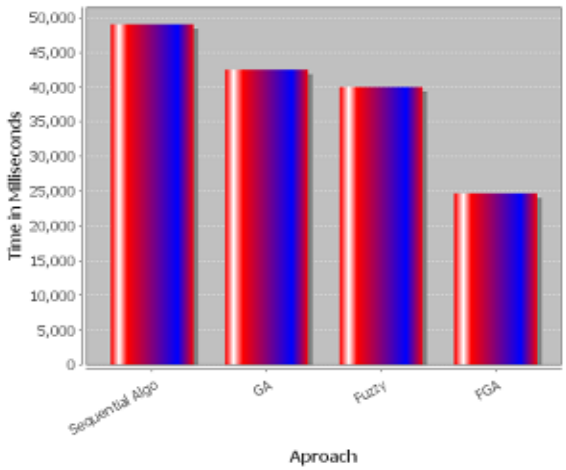

Fig. 2 Comparison between the proposed system and existing systems in terms of time

Figure 3 shows a comparison between the proposed system and existing systems in terms of memory allocation on the system.

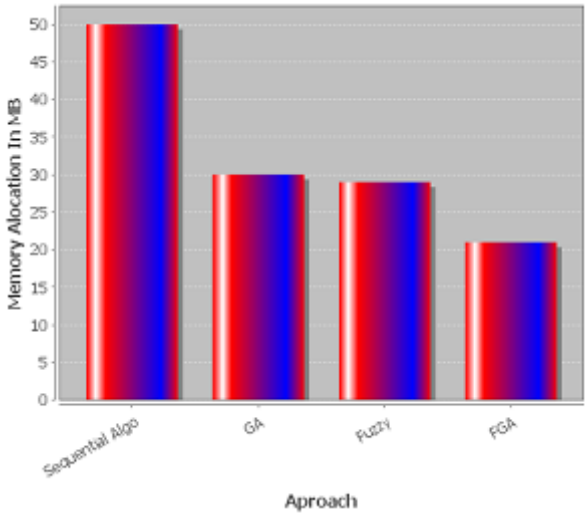

Fig. 3 Comparison between the proposed system and existing systems in terms of memory

VI. CONCLUSIONS

In this paper, a fuzzy genetic algorithm is proposed for dealing with the intrusion detection problem considering KDD99 dataset. Results are compared with the existing system which uses sequential algorithm for intrusion detection. The results show that the accuracy of detection rate of the proposed system for DoS, probe, Remote to User Attacks (R2l) and User to Root attack (U2r) are more compared to the existing systems. The time required for the training and testing of the dataset using the proposed system is less compared to the existing systems and memory allocation also requires less space for proposed system than existing systems.

REFERENCES

[1] Bridges S.M and Vaughn R.B, "*Fuzzy Data Mining and Genetic Algorithms Applied to Intrusion Detection*," Proceedings of 12th Annual Canadian Information Technology Security Symposium, pp. 109-122, 2000.

[2] Bridges, Susan and Rayford B. Vaughn. 2000. "*Intrusion Detection via Fuzzy Data Mining*," In Proceedings of 12th Annual Canadian Information Technology Security Symposium, pp. 109-122. Ottawa, Canada

[3] Crosbie, Mark, and Gene Spafford. 1995. "Applying Genetic Programming to Intrusion Detection." In Proceeding of 1995 AAAI Fall Symposium on Genetic Programming, pp. 1-8. Cambridge, Massachusetts.

[4] J. Gomez and E. León, "A fuzzy set/rule distance for evolving fuzzy anomaly detectors," IEEE International Conference on Fuzzy Systems ART. No. 1682017, pp. 2286-2292.

[5] N. Ngamwitthayanon and N. Wattanapongsakorn, "Fuzzy- ART in network anomaly detection with feature-reduction dataset," The 7th International Conference on Networked Computing, INC2011, Art. No. 6058956, pp. 116-121.

[6] P. Jongsuebsuk+, N. Wattanapongsakorn+, C. Charnsripinyo "Real-Time Intrusion Detection with Fuzzy Genetic Algorithm." ©2013 IEEE.

[7] T.P. Fries, "A fuzzy-Genetic approach to network intrusion detection," GECCO'08: The 10th Annual Conference on Genetic and Evolutionary Computation, 2008, pp. 2141-2146.N.

[8] W. Li, "Using Genetic Algorithm for Network Intrusion Detection." "A Genetic Algorithm Approach to Network Intrusion Detection" SANS. Institute, USA, 2004

