Host based Anomaly Detection using Fuzzy Genetic Approach (FGA)

Harjinder Kaur
M.E (CSE) Student, Punjabi University
Regional Centre of IT & Management, Mohali,
Punjab, India

Nivit Gill
Assistant Professor (CS), Punjabi University
Regional Centre of IT & Management, Mohali,
Punjab, India

ABSTRACT

Intrusion is a fast growing security threat to the computers which fails the security of the system. The researchers have proposed number of techniques such as firewall, encryption etc. to prevent such penetration and protect the systems. With all these measures also, the intruders managed to penetrate the computers. Intrusion detection systems (IDS) monitor the resources of the computer, detect the malicious/suspicious activity either on a single machine or on the network which is different from the legitimate user activity and send the reports of such activity. The paper proposes to detect anomalous user behavior on a single machine based on the system log files using fuzzy logic and genetic algorithms.

KEYWORDS

Intrusion, Signature/Misuse, Anomaly, Fuzzy Logic and Genetic Algorithm.

1. INTRODUCTION

Computers have revolutionized the IT industry, by changing the traditional way of working. That is, all the manual work has been taken over by the machines which save lot of time and effort. This raised the security concern which can either be violated by the internal threats like viruses, worms etc. or by the illegitimate users. Any activity which violates the security policy of the system refers to as intrusion in the information system and intrusion detection is a process used to identify such intrusions [1]. Intrusion detection system observes the activities on the single host or a network and if any unusual activity is detected then the intrusion detection system reports such activity.

In this paper, the user behavior of a single machine/host has been analyzed from the log files of the system. The normal behavior of the user is encoded into fuzzy rules and based on these fuzzy rules the behavior of the user is studied and is optimized with the help of genetic algorithms. The organization of the paper is: section 2 reviews the literature related to intrusion detection system, section 3 presents the overview of IDS, sections 4 and 5 provide the basic concepts of fuzzy logic and genetic algorithms respectively. The proposed fuzzy genetic approach to host based anomaly detection is explained in section 6. The simulation results are given in section 7 followed by conclusion in section 8.

2. RELATED WORK

Various authors have proposed a number of techniques for the detection and prevention of the intrusion in the system**.** S. Mallissery et al. [1] and A. S. Ashoor et al. [2] review the various categories of intrusion detection techniques. The authors have categorized the intrusion detection techniques on the basis of data source and on the model of intrusions. V. Jyothsna et al. [3] and P. G. Teodoro et al. [4], classify the anomaly detection techniques into statistical, cognition and machine learning based techniques. However with all these techniques there is a need of hybrid intrusion detection which can detect static as well as dynamic attacks i.e can detect attacks on host as well as on network. B. Shanmugam et al. [5] proposed a hybrid IDS which is more accurate, low in false alarms, intelligent by using fuzzy mechanisms and not easily deceived by small variation. But the system crashed, as it could not withstand the traffic for more than three weeks without restarting. The author M. Hassan [6] reviewed some of current and past intrusion detection technologies, which includes fuzzy logic and genetic algorithms. Author also proposed a work that implements the Genetic Algorithm (GA) and fuzzy logic using the KDD 99 dataset of network intrusion detection. GA efficiently identifies various types of network intrusions and fuzzy logic determines false alarm rate by which intrusive activities can be minimized.

L. Ying et al. [7] proposed a technique of host based IDS which uses log file analysis technology and back propagation (BP) neural network technology. Where Log file analysis is used in misuse detection, and BP neural network is used in anomaly detection. By combining these two detection technologies, the HIDS implemented by the authors showed effective improvement in the efficiency and accuracy of intrusion detection. A novel approach for design of fuzzy rules using genetic algorithm is given by S.V. Wong et al. [8]. In this paper, Genetic optimization library (GOL) is developed and implemented to get the set of fuzzy rules, where the developed GOL replaced the process of trial and error to obtain the better combination of fuzzy rules.

A fitness calculation has been employed to handle the maximization and minimization problem. R. Borgohain [12] and S. Owais [14] et al. explain the fuzzy logic and genetic algorithms in detail and discussed various rules and parameters of these techniques. N. Benamrane et al. [15] proposed an approach for detection and specification of anomalies present in medical images using fuzzy neural networks and genetic algorithms.

In this paper, hybrid fuzzy genetic approach has been applied to host based intrusion detection on the dataset of system log files. The hybrid of fuzzy logic and genetic algorithm has been applied by the author M. Hassan for the network based intrusion detection [6].

3. INTRUSION DETECTION SYSTEMS (IDS)

Intrusion detection systems (IDS) are used to detect intrusion in the computer. Generally the IDS are categorized on the basis of data source and on the basis of model of intrusions, as illustrated in Figure 1. The data source based classification is further classified into two types: Host based IDS (HIDS) and network based IDS. The host based IDS detect the intrusion in the single machine only whereas Network based IDS (NIDS), detects the intrusion on a network. These two techniques can be combined together to form a Hybrid IDS [1]. The second category of IDS based on model of intrusion is classified into Signature/Misuse based intrusion detection systems and anomaly based intrusion detection systems. The signature based IDS depends on the receiving of regular updates of patterns and comparing those patterns with the predefined signature and able to detect known previous threats and the Anomaly based intrusion detection, which used the rules and the heuristics for defining the behavior as normal or anomalous rather than the predefined signature, thus it can detect previous unknown threats [2]. Anomaly based detection can be employed on a host or a network.

Anomaly based detection can be classified into Statistical based detection systems, Cognition based and Machine learning based systems [4]. Statistical based detection systems capture the network traffic activity and generate a profile based on the stochastic behavior of the network. This profile is based on metrics such as the traffic rate, the number of packets for each protocol, the rate of connections, the number of different IP addresses, etc. [4]. Cognition-Based (also called knowledge-based or expert systems) detection systems work on a set of predefined rules, classes and attributes identified from training data, set of classification rules, parameters and procedures inferred. Machine learning based systems are used where no information is available regarding ground truth and they use of a small subset of data points are used to estimate the unknown attributes of test points [3]. Many machine learning based detection techniques are used for the host as well as the network based anomaly detection. Bayesian Networks are used to encode probabilistic relationships among the variables of interest and ability to incorporate prior knowledge and data [4], Whereas Neural Networks generalize from noisy, limited and incomplete data and have the potential to recognize future unseen patterns [5]. Fuzzy Logic is employed where reasoning is approximate rather than precise [4],[5]. Genetic Algorithms and evolutionary algorithms are capable of deriving rules and selecting optimal parameters by use of the fitness function [6], [11], [13].

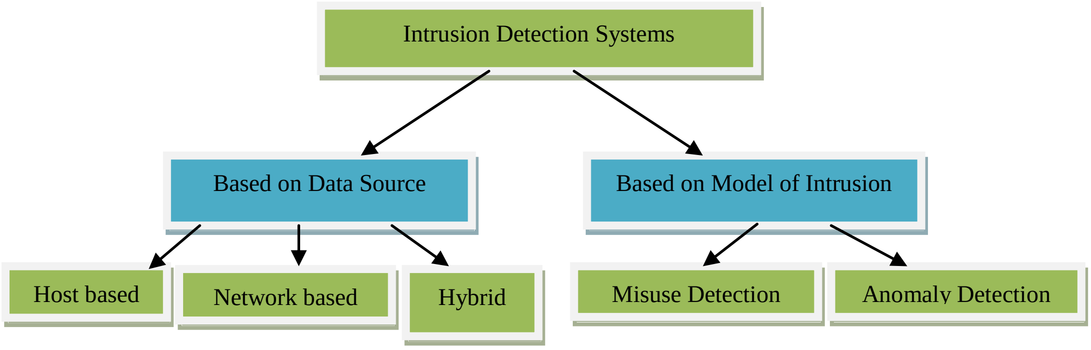


Figure 1: Types of intrusion detection systems [2]

4. FUZZY LOGIC

Fuzzy logic is derived from fuzzy set theory where reasoning is approximate rather than precisely inferred from classical predicate logic. The theory of fuzzy sets, which is based on fuzzy logic was introduced by Lofti Zadeh in 1965 as the mathematical way to represent vagueness in linguistics and can be considered as a generalization of the classical set theory [9]. In classical theory, an element either belongs to the set or not, while in fuzzy logic the fuzzy sets eliminate the sharp boundary that divides members from non-members in a group. The degree of membership is defined using a generalized function known as the membership function as below:

$$A(u):U \to \quad [0,1]$$

Where U is a universe and A is a fuzzy subset of U. The values of membership functions are real numbers between [0, 1], where 0 means the object is not a member of the set and 1 means completely belongs to the entire set. Each value of the function is called membership degree [9].

Fuzzy systems mainly contain three components: Fuzzifier, Inference Engine and Defuzzifier. Fuzzifier converts the crisp inputs to linguistic variables using the membership functions. The second component Inference Engine used the Fuzzy rules of type If-Then that converts the fuzzy inputs to the fuzzy outputs. Various inference systems are Mamdani Fuzzy models, Sugeno Fuzzy Models and Tsukamoto Fuzzy Model. Defuzzifier gives the crisp output by converting the fuzzy output of the inference engine using membership functions [10].

5. GENETIC ALGORITHM

Genetic algorithm (GA) is now a very popular and widely used tool to solve the optimization problems. It has been used effectively to find optimal solutions for the variety of problems. Genetic algorithms are based on the theory of biological natural selection and natural genetics. Error! Reference source not found. This is achieved by using the fitness function where fitness leads to the efficiency and higher the value of fitness, higher is the efficiency. A high fitness value is the characteristics of a good chromosome. From the initial population the genetic algorithm finds the solution to the optimization problem by use of five basic operations which are used to evolve the population from one generation to the next generation. These operations are selection, reproduction, crossover, mutation and fitness [11] [12].

Selection is the process of reproduction to select individuals to create the next generation. This is driven by fitness functions that produce higher fitness. Reproduction and crossover involves the crossing of the individual's chromosomes to produce the chromosome of the offsprings. It is a random process which occurs when chromosomes of two parents mate with one another and both chromosomes are dissected at the same predefined crossover point [11]. Mutation occurs in a chromosome with a very small probability. When it occurs in a chromosome then a random bit in binary chromosome is inverted. Mutations are too much important as they explored the solutions, that may have not previously occurred in optimise search space [14]. The Fitness or a cost of a chromosome is used to determine how ‘good’ a chromosome is, and a high fitness indicates a ‘good’ chromosome. Fitness functions are applied to each new individual in genetic algorithms. It must be scaled so that to return a non-negative value. Higher performing individuals are always obtained from the higher fitness values [13].

6. PROPOSED WORK OF THE RESEARCH

In this paper, host based anomaly detection has been done using fuzzy genetic approach in which the system log files are analyzed to detect the user behavior of the system. Earlier combination of fuzzy logic and genetic algorithms has been used successfully for network based intrusion detection [6]. The log files record the behavior of computer system and aim at recording the action of operating system, applications, and user behaviors. Error! Reference source not found. The proposed work has been divided into two phases: phase 1 uses the fuzzy rules that are constructed from the user behaviour and in phase 2 the fitness function is applied on the fuzzy output of the phase 1 to obtain the optimized result. The proposed layout of this research work is shown in Figure 2:

Phase 1 used the fuzzy approach on the host for the anomaly detection where user behavior is extracted from the data set and then by applying the fuzzy rules and the membership functions results are obtained. The fuzzy approach is applied using the three components; Input, Output and the rules explained as follows:

- **Input:** There are two inputs to the fuzzy toolbox taken from the user behavior that is the set of events and the timing information. Here the set of events are used to describe the activities observed in the system and the timing information records the particular time of the activity such as login and the logout timings.
- **Output:** From the system log it has been observed that the output of the activity either signifies the information or a warning**.**
- **Rules:** From these set of inputs and outputs, the fuzzy rules have been designed.

Phase 2 used the genetic approach for the optimization, in which a fitness function is applied on the fuzzy output and the optimized result is found using the best fitness of the algorithm.

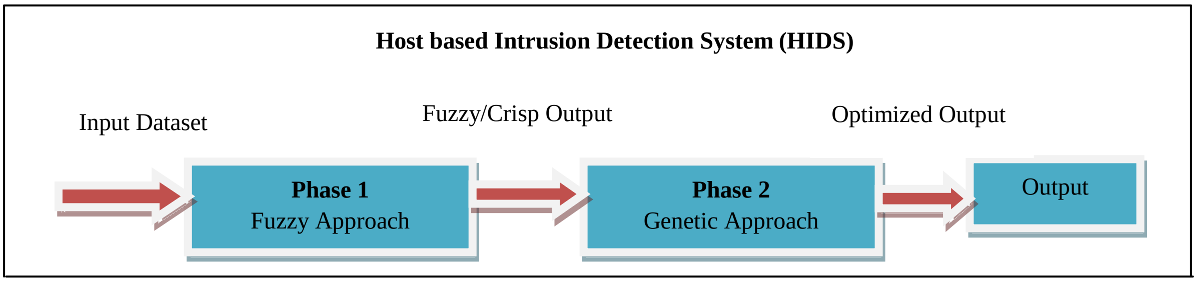
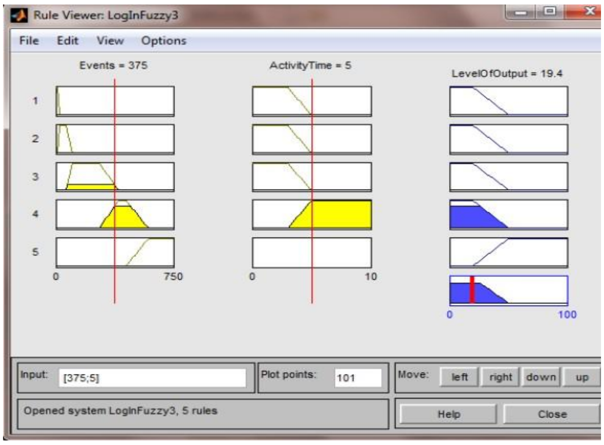


Figure 2: Proposed Work of the Research

7. SIMULATION & RESULTS

The proposed work has been implemented using MATLAB as it provides toolboxes for both fuzzy logic and genetic algorithms. System log files of seven days have been studied and analyzed to get the set of events and timing information. The fuzzy output is the level of output, which is either the information or a warning. Based on this data, various rules are designed in the MATLAB fuzzy toolbox as shown in Figure 3:



Input: [375;5]

Plot points: 101

Opened system LogInFuzzy3, 5 rules

Figure 3: MATLAB Fuzzy Rule Viewer showing the membership of the inputs and the outputs

The genetic fitness function has been designed by getting the results of the fuzzy approach. Ten independent runs are carried out to verify genetic optimization's performance of fuzzy rules design as shown in Table 1. The number of generations is set to 100. All the runs share the same genetic parameters, such as mutation probability, the number of individuals in the population, but all have different random seeds. The random seeds are used to generate random numbers.

After analysis of this result, it has been found that since genetic algorithm uses random number generators, the algorithm returns slightly different results each time it runs. The performance of the genetic optimization firstly reduced when finer solution evolves. Smaller improvement degree is experienced when whole of the population evolves further. This illustrates that genetic optimization is performing excellent initially, and then performance gets deteriorates from generation to generation till it reaches to its optimum point. The best fitness or the optimum point is obtained at 65 iterations.

Table 1: Result of genetic algorithm in HIDS for the Optimization

S.no.	Current Iteration	Best Fitness
1	78	11.5255
2	77	11.5156
3	66	11.5212
4	65	11.5726
5	78	11.4934
6	69	11.5288
7	67	11.5083
8	65	11.5044
9	64	11.5425
10	82	11.5199

8. CONCLUSION

In recent years various intrusion detection systems have been proposed to protect the systems against the various types of attacks: static or dynamic. These systems used different methods and techniques in an attempt to deliver better efficient intrusion detection systems. In this paper, fuzzy logic and genetic algorithms have been used for host based anomaly detection. Fuzzy logic is quiet popular due to its approximate reasoning and genetic algorithms (GA) are successful for finding optimal solution effectively. In the proposed work, log files are used to detect the user behavior of the system and fuzzy rules are derived from the information obtained. Genetic Algorithms are applied on the fuzzy output and optimized results are obtained. The experimental results of the simulation of the proposed work show impressive results where the best fitness value is the ideal fitness value and smaller improvement degree is experienced till it reaches the optimum point. However GA is a randomization search method so it returns slightly different results each time it runs. So by using hybrid Fuzzy Genetic Approach (FGA), optimized results of anomaly detection are obtained over the user behaviour of the single machine/host.

9. REFERENCES

[1] Sanoop Mallissery, Jeewan Prabhu, andRaghavendra Ganiga, "Survey on IntrusionDetection Methods," in proc. of 3rd Int. Conf. onAdvances in Recent Technologies in Communication and Computing, Bangalore, 2011,pp. 224-228.

[2] Asmaa Shaker Ashoor and Sharad Gore, "IntrusionDetection System: Case study," in InternationalConference on Advanced Materials Engineering, vol. 15, Singapore, 2011, pp. 6-9.

[3] V Jyothsna, V V Ramaprasad, and K MunivaraPrasad, "A Review of Anomaly based Intrusion Detection Systemss," International Journal of Computer Applications, vol. 28, no. 7, pp. 26-35, August 2011.

[4] P Garcia Teodoro, J Diaz Verdejo, G Macia Fernandez, and E Vazquez, "Anomaly based network intrusion detection: Techniques, Systems and Challenges," International Jouirnal ofComputers and Security, vol. 28, no. 1, pp. 18-28, February-March 2009.

[5] Bharanidharan Shanmugam and Norbrik BashahIdris, "Hybrid Intrusion Detection Systems(HIDS) using Fuzzy Logic," in Intrusion Detection Systems,Dr. Pawel Skrobanek, Ed. Croatia, Europe: InTech,2011, ch. 8, pp. 135-155.

[6] Mostaque Md. Morshedur Hassan, "Current Studies on Intrusion Detection System, Genetic Algorithm and Fuzzy Logic," International Journal of Distributed and Parallel Systems (IJDPS), vol. 4, no. 2, pp. 35-47, March 2013.

[7] Lin Ying, Zhang Yan, and Ou Yang Jia, "The Design and Implementation of Host-based IntrusionDetection System," in 3rd International Symp. On Intelligent Information Technology and Security Informatics, Jinggangshan, 2010, pp. 595-598.

[8] S. V. Wong and A. M.S Hamouda, "Optimization of fuzzy rules design using genetic algorithm," Advances in Engineering Software, vol. 31, no. 4, pp. 251-262, April 2000.

[9] Anne Magaly, "Combining Neural Networks and Fuzzy Logic for Applications in Character Recognition," University of Kent, Canterbury, PhD Thesis 2001.

[10] Mathworks, Fuzzy Logic Toolbox User's Guide. Natick, United States: Mathworks Inc., 2012. [11]Mathworks, Genetic Algorithm and Direct search toolbox For use with Matlab. Natick, United States: Mathworks Inc., version 1, 2004.

[12] Rajdeep Borgohain, "FuGeIDS: Fuzzy Genetic paradigms in Intrusion Detection Systems," International Journal of Advanced Networking and Applications, vol. 3, no. 6, pp. 1409-1415, 2012.

[13] Adel Nadjaran Toosi, Mohsen Kahani, “A new approach to intrusion detection based on an evolutionary soft computing model using neurofuzzy classifiers,” Computer Communications, vol. 30, no. 10, pp. 2201–2212, 2007.

[14] Suhail Owais, Vaclav Snasel, Pavel Kromer, Ajith Abraham, "Survey: Using Genetic Algorithm Approach in Intrusion Detection Systems Techniques," 7th Computer Information Systems and Industrial Management Applications, Ostrava, pp. 300-307, 2008.

[15] N.Benamrane, A. Aribi, L.Kraoula, "Fuzzy Neural Networks and Genetic Algorithms for Medical Images Interpretation," Conference on Geometric Modeling and Imaging--New Trends, London, pp.259-264, 2006.

[16] Jian Xu, Jing You, and Fengyu Liu, "A Fuzzy Rules based Approach for Performance Anomaly Detection," proc. of IEEE Networking, Sensing and Control, pp. 44-48, 2005.