Intrusion Detection System using Genetic Algorithm: A Survey

Vaibhavi Pandya[1] Gargi Chauhan[2]
[1]M.Tech. Student [2]Assistant Professor
[1,2]Department of Information Technology (System & Network Security)
[1,2]Sardar Vallabhbhai Patel Institute of Technology Vasad, India

Abstract— Today, It is very important to provide a high level security to protect highly sensitive and private information. Intrusion Detection System is an essential technology in Network Security. Nowadays researchers have interested on intrusion detection system using Data mining techniques as an artful skill. So my aim is to use IDS system and improve the performance of the IDS. IDS is a software or hardware device that deals with attacks by collecting information from a variety of system and network sources, then analyzing symptoms of security problems. The main focus of Intrusion detection and prevention systems (IDPS) is to identify the possible incidents, logging information about them and in report attempts. But sometimes IDS is not able to detect some new attacks and vulnerabilities. The main aim of this paper is to provide the knowledge of the IDS and also gives the survey of genetic algorithm for IDS. So this paper provides basic knowledge of IDS system as well as how to use genetic algorithm for IDS and improve the future research.

Key words: Genetic Algorithm, IDS, Fitness Function, NIDS, HIDS, Detection Rate, False Positive Rate

I. INTRODUCTION

Intrusion Detection System (IDS) is as the part of the security and security is the essential part for any online network now days hence it is in picture.

An intrusion is "a collection of actions that aspire to understand the confidentiality, integrity or accessibility of different resource". Intrusion can also be define as "a collection of actions conceive to acquire unauthorized resources, misuse rights, cause complete systems and networks crashed, decrease running potency, or deny services". Thus, IDS may be a system to observe events in computers or networks and analyses and monitoring the systems integrity and confidentiality. Genetic Algorithm is a search heuristic that gives a useful solution to search and optimization problems. These algorithms convert the problem in a specific domain into a model by using a chromosome-like data structure and evolve the chromosomes using selection, recombination, and mutation operators. Genetic algorithm use as problem solving strategy and provide optimal solution of the problem. Genetic algorithm works on the Darwinian principle of reproduction. It is transform set of individual objects, which each associated fitness value into new generation of population and then apply crossover and mutation function.

II. TERMINOLOGIES RELATED TO IDS

A. *Characteristics of Good ids*

IDS are pattern recognition systems similar to anti-virus software. The difference is that anti-virus software inspects the files on your computer where IDS looks for attacks on the network. IDS must run continually without human supervision. IDS must be also fault tolerant means it must survive a system crash. IDS also impose minimum overhead on the system.

B. *Vulnerability of ids*

The system can collect a large number of alerts in a day, so that overloading of the work. It is common for the number of real attacks to be far below the number of false alarms. Number of real attacks is often so far below the number of false-alarms that the real attacks are often missed and ignored. NIDS cannot properly protect high-speed networks. Tasks like analyzing and filtering has to be done manually.

III. OVERVIEW OF IDS FOR GENETIC ALGORITHM

Here I referred some papers for the survey from which I give brief overview of all the papers.

A. *Using Genetic Algorithm Approach in Intrusion Detection Systems Techniques: survey*

Computers are devices that store, process, and retrieve data. Data is invaluable resource for every company. The most important requirements for handling data or securing data are availability, integrity and confidentiality. This three are the main requirement for any data to keep it secure. In recent years, there are increase the amount of data that available on the Internet. Therefore, the hackers and intruders had made many successful attempts to bring down high-profile company networks. As reported by the Computer Emergency Response Team/Coordination Center (CERT/CC), the number of computer attacks has increased exponentially in the past few years from 1990 to2003 as shown in figure.[1]

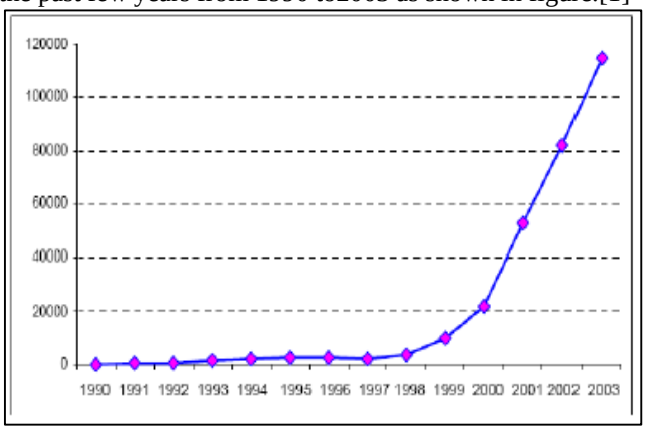

Fig. 1: Growth rate of cyber incidents

Several security functions provide IDS to us, The three fundamental security function components that are provided by IDS likes (i) *Information sources which is* monitor computers or network for an unauthorized entrance; activity; file modification. It also gives the information about intruders take place or not in the system. (ii) Analysis which is detect any abnormal activity. There are two approaches for the analysis like misuse detection and anomaly detection. (iii) Response which is set of activity that the system perform when any intruder takes place.

There are three different components of IDS: (i) Sensors or Agent which is generate security events. (ii) Console is the program that provides the interface to the administrator or IDS user. (iii) Engine which is records events logged by the sensors in a database and use a system of rules to generate alerts.

GA evolves the population of chromosomes (individuals) as the process of natural selection. It generates new chromosome which is also known as offspring, during the process. GA process uses three types of main function or genetic operators which are selection operator, crossover operator and mutation operator. Using these operators evaluate chromosome using the fitness function. There are many researchers that used evolutionary algorithms and especially GAs in IDS to detect malicious intrusion from normal use. In this paper some of the most important researches on using GAs in IDS will be discussed.

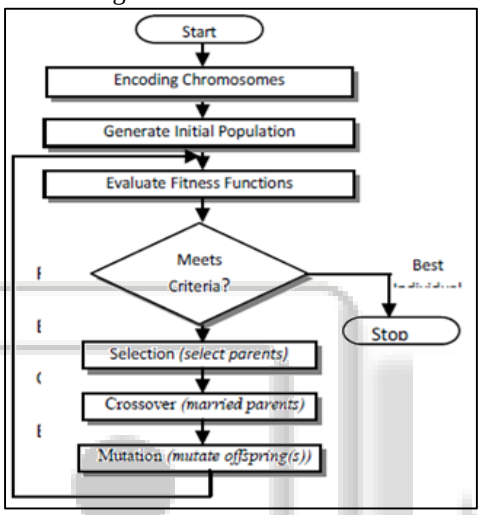

Fig. 2: Genetic algorithm process

Chitturs paper in 2001 presents a novel approach to detect the malicious intrusions (*hacks*) by using a complex artificial intelligence method known as GA applied to IDS. The researcher applies GA to learn how to detect malicious intrusions and separate then from normal use. Using GA result gives us ; the best fitness value was very closely to the ideal fitness value of 1. The system able to detect about 97% of attacks and 0.69% of normal connections were incorrectly classified as attacks.

Jiu-Ling Zhao et al. in 2005 represented about IDS using GA that, Misuse detection system and anomaly detection system encode an expert's knowledge of known patterns of attack and system vulnerabilities as *if-then* rules. They also used two methods for cluster analysis, one is hierarchical and another one is K-means. They conclude that only about 0.71% of normal connections were classified as attacks; also have very low false positive rate. The GA was successfully applied what it had learned to a real-world test case. And clustering GAs are promising method for the detection of malicious intrusions into computer systems.

Pedro A. Diaz-Gomez et al. in 2006 used the evolution process set of possible solutions were generated randomly. In that they evaluate each chromosomes using fitness function. They also use single point crossover and single point mutation. In their research they perform GA for offline Intrusion Detection. As a result they test the system by implementing different formulas for fitness function. They found that there are no false positive and the number of false negative decreases dramatically. GA engines an appealing tool in the search for intrusions in audit trail files.

Ren Hui Gong et al. in 2005 choose the approach to network misuse detection. The result shows that the GA approach is very effective and also have flexibility to detect the intruder and also classify them. In this approach there is good detection rate and depending on the selection of fitness function weight values, the generated rules can be used to either generally detect network intrusions or precisely classify the types of intrusions.

B. Survey on Intrusion Detection Methods

Intrusion detection is very important aspects of protecting the cyber infrastructure from terrorist attack or from hackers. There are many Intrusion prevention technique such as firewall, filtering router policies fails to stop much type of attacks. Intrusion still happens and so they must be detected. Therefore Intrusion detection systems are becoming an important part of your computer system for more security. There are two types of IDs one is Host Based IDS and another one is Network Based IDS.[2]

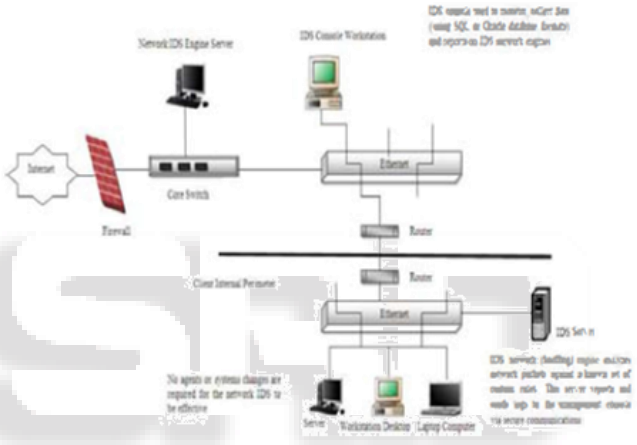

Fig. 3: Network Based IDS

In network based IDS they monitor hole the network system using only one IDS system. NBIDS are valuable if they placed just outside the firewalls. They monitor hole traffic of the network packets and find if any abnormal activity is there in network or not.

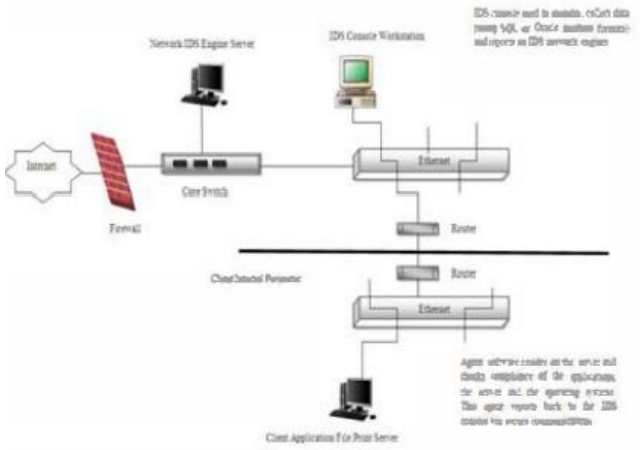

Fig. 4: Host Based IDS

In Host Based IDS software agents are installed on each of the computer hosts of the network to monitor the events occurring within that host only. It performs log analysis, file integrity checking, policy monitoring, real-time alerting and active response. HIDS overcome the problems

incurred in Network based IDS technology of securing individual hosts in the network.

There are different methods for IDS. In that one is pattern matching technique which is based on looking for a fixed sequence of bytes in a single packet. It is an approach that is fairly rigid but simple. The structure of a signature based on the simple pattern-matching approach might be as follows if the packet is IPv4 and TCP and the destination port is 2222 and the payload contains the string "foo," fire an alarm. This example of a pattern match, of course, is a very simple one, but the variations from this point are also simplistic to employ. Another method is state full pattern matching in which systems that perform this type of signature analysis must consider arrival order of packets in a TCP stream and should handle matching patterns across packet boundaries. Another one is protocol decode based analysis in which class of signature is implemented by decoding the various elements in the same manner as the client or server in the conversation would. One more method is anomaly based analysis which is typically geared to looking for network traffic that deviates from what is seen "normally." The biggest problem with this methodology is to first define what "normal" is.

C. IGIDS: Intelligent Intrusion Detection System Using Genetic Algorithm

In this paper here they present a genetic algorithm based network intrusion detection system named IGIDS, where the genetic algorithm is used for pruning best individuals in the rule set database. Here the process makes the decision faster as the search space of the resulting rule set is much compact when compared to the original data set, therefore this makes IDS faster and intelligent. They also generate possible intrusions which forms the basis for detecting intrusions on the network traffic. Therefore the result gives the high detection rate and low false positive alarm rate.[3]

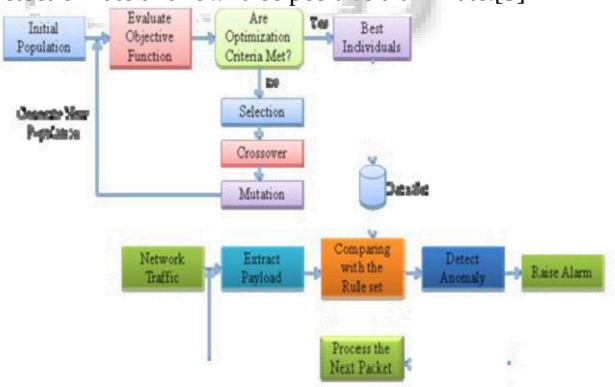

Fig. 5: Detailed System Architecture for IGIDS

Here in proposed work they use payload base system. For build the profile for each port monitored, they use destination address and service port numbers only this two headers they use. They first use DARPA dataset to convert important information in to the rules. After that rules are used for reproduction of the new rules and calculate the fitness value. Using this fitness value it is decided to the new rules intrusive or not. The drawback of this system is IDS will have to be trained for every new type of application and a lot of legitimate traffic may be classified as an attack. In this system the decision tree is used for calculating the information gained from each attribute which is then use to calculate the weight of the attributes. These weights will be used to decide the priority of the attributes.

Here Researchers have successfully evolved their rule sets to detect new intrusions. Therefore the system can be deployed with any intrusion detection system for evolving the rule set and detecting intrusions from the incoming traffic in the system that uses it. Here they include, the future work can be aimed at integrating with an Intrusion Detection System such that apart from existing intrusions even new intrusions can be detected.

D. Implementing Rule based Genetic Algorithm as a Solution for Intrusion Detection System

In this paper the authors are going to present Genetic Algorithm to identify various harmful/attack type of connections using different features in network connections and each rule set identifies a specific type of attacks. Here the characters of attacks like Smurf, Warezmaster, Saint, Mail bomb, multihop, IP sweep, snmpguess, buffer-overflow were summarized through the KDD99 data set and the effectiveness and robustness of the approach has been proved. Using these rules it is helpful for working with high quality accuracy for detecting the denial of service and probe attack and also with appreciable accuracy for identifying the U2R and R2L connections. [4]

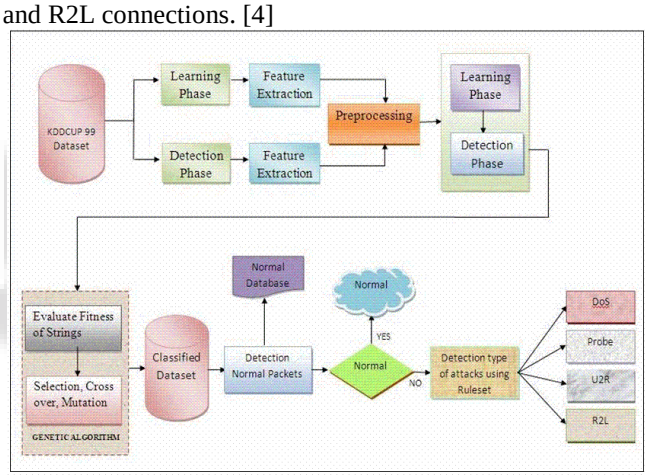

Fig. 6: The simple Architecture of the proposed model

Here in the architecture there are two phase, one is learning phase which contain rule set for detecting intruders using network audit data and second phase contain best rule set with highest fitness value used for detecting intruders in the Internet world. In this system every rule follows the construction based on IF-THEN format. In preprocessing, the symbolic features are converted into binary form and normalize the data and in the detection phase they have applied Genetic Algorithm on selected features dataset and find fitness for each rule using the following fitness function.

$$\text{Fitness} = f(x) / f(\text{sum}) \quad (3.1)$$

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

Here they use rank selection. First individuals are sorted and ranked based on their fitness value.

$$Ps(i) = r(i) / \text{rsum} \quad (3.2)$$

Where Ps (i) is probability of selection individual r(i) is rank of individuals rsum is sum of all fitness values. They collect the classified dataset from the Genetic Algorithm and rules applied to detect the errors.

S. No	Attack Name	Detection Percentage
1	Mailbomb	87%
2	Warezmaster	98%
3	Multihop	73%
4	Smurf	73%
5	Snmpguess	99.87%
6	buffer_overflow	65%
7	Saint	77%
8	Ipsweep	98%
Average Success Rate		83.85%

Table 1: Result table of attack type and detection rate

Here authors use genetic algorithm for detecting DOS, Probe, u2R and R2L attacks from KDD99 dataset. The result provides good detection rate 83.85%. This system also provides flexibility for usage in different application areas with proper attack taxonomy. In future try to improve result of whole system and reduce the complexity of the system and also reduce the training time using more reduction techniques.

E. Effect of Change in Rate of Genetic Algorithm Operator on Composition of Signatures for Misuse Intrusion Detection System

In this paper authors conclude about the effect of change in rate of genetic algorithm fitness value in composition of signatures for misuse intrusion detection system. Here proposed work uses a set of classification rules which are generated from the predefined intruders.[5]

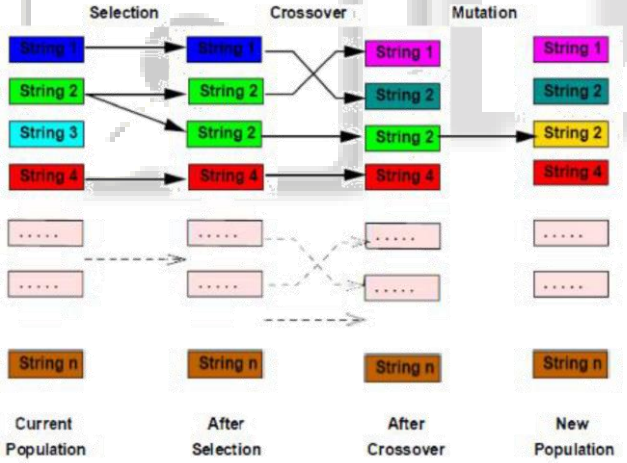

Fig. 7: Working principle of the genetic algorithm

Here they first choose random generated population and then after calculate total number of the records. After that calculate the fitness of all record and find the best fittest chromosomes. Here they take different crossover rate and mutation rate is constant. From this experiment conclude that, if crossover performed is not well then there is not sufficient sharing of genes. If we crossover too much, good segments of individuals get split up a lot. This allows some individuals with high fitness's to be copied directly to the next population.

F. Intrusion Detection System Using Genetic Algorithm

In this paper author uses genetic algorithm approach with improve population and selection operation. Here the main aim of this paper is the optimization of the number of signatures in the audit file to minimize the search time and increase the detection rate of attacks. In this experiment they use NSL-KDD99 dataset and choose seven attributes from audit data.[6]

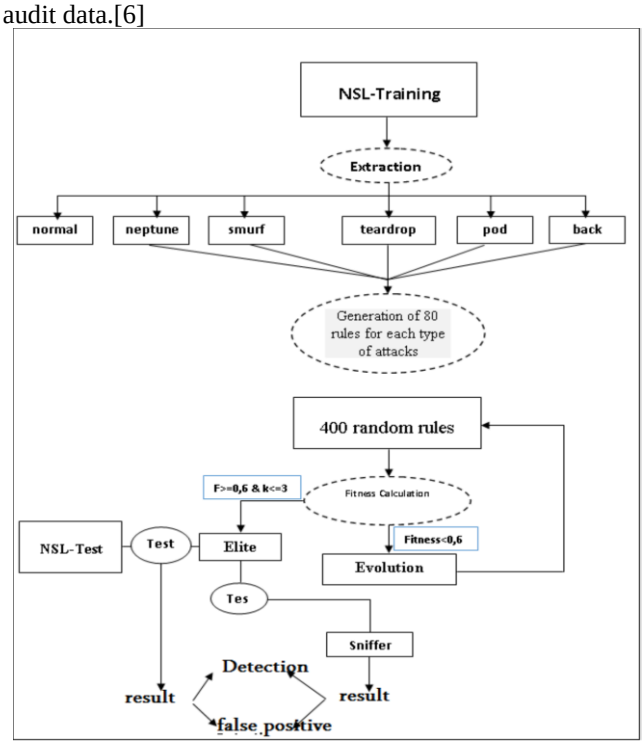

Fig 8: Architecture of applying GA into IDS

In the proposed work first they collect enough historical data which contain both normal as well as abnormal behavior of network connection. The dataset generated on random basis here they uses 80 rules for each type of attack. Then after it is analyzed by the network sniffer and result fed in to the genetic algorithm. Then apply GA and rule sets are generated which is store in to the database that used for IDS.

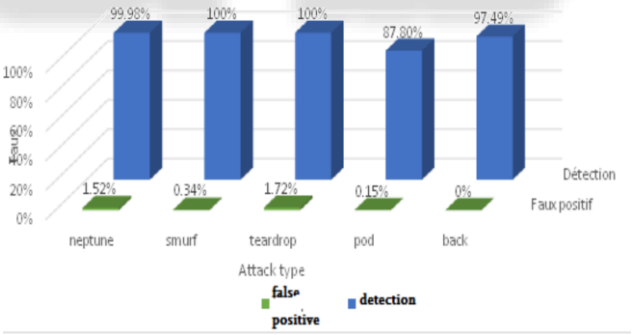

Fig. 9: Result of detection for each attack

In this work they are try to improve the search time in audit data without losing the performance of the system and they get satisfactory result for IDS based on genetic algorithm. In the result they produced very high detection rate (99%) and low false positive rate (3%).

G. Genetic algorithm with Different Feature Selection Method for Intrusion Detection

In this system researchers proposed different feature selection methods like information gain, mutual correlation, and cardinality of features. In this model they implement several steps like (a) dataset preparation (b) prepare train and test set (c) feature selection techniques (d) implementation of genetic algorithm (e) prediction. [7]

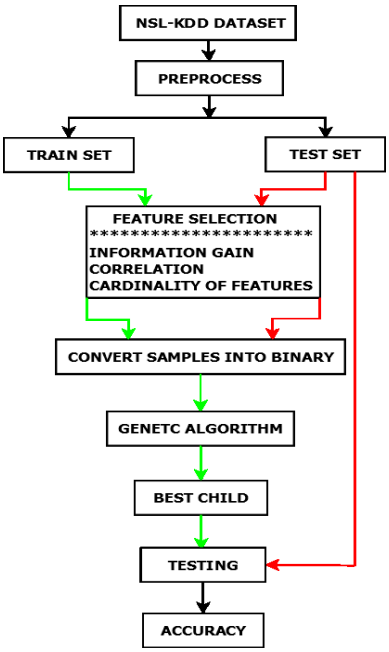

Fig. 10: Proposed Architecture for Genetic Algorithm

The main aim of this model is to detect the intrusions with minimal number of features using genetic algorithm. They evaluate the different features which are obtained from the different feature selection technique then after they implement genetic algorithm on the dataset. Here they use min-max normalization for the creating homogeneity between the features. After normalization they have converted normalized instant in to binary values and algorithm implemented on to these binary values. At last they produced best child using genetic algorithm. Here different population sizes are used ex. 50, 100 … 400. Algorithm runs on 100 iterations. Here they use binary encoding in which each chromosomes represented in to 0's and 1's. After the normalization, the value between 0.0-0.25 represent is represented as "00", 0.25-0.50 is represented as "01", 0.50-0.75 represented as "10", finally 0.75-1.0 is "11". The classes are denoted as 00,01,10,11 (Normal, Neptune, Satan, and Smurf). In the GA operator they use two point crossover , bit flip mutation and random selection. Here fitness is defined as

$$\text{Fitness (x)} = \frac{a}{A} \qquad (3.3)$$

Where, a is the number of samples that exactly matches an individuals and *A* is the total number of normal samples. Minkowski distance measure is used to calculate the distance between a child and parents.

$$d(X,Y) = \left(\sum_{i=0}^{n}(xi - yi)^p\right)^{1/p} \qquad (3.4)$$

Classification performance of different kind of features measured in terms of accuracy. Accuracy defines as a ratio of total number of instances correctly classified to the total number of samples in the data set.

$$\text{Accuracy} = \frac{TP+TN}{(TP+FN+FP+TN)} \qquad (3.5)$$

Where TP is the number of normal samples classified as normal, FN is the misclassified instances of the normal class, TN is correctly classified instances of abnormal class, and FP is the number misclassified instance of abnormal class.

In the result Information gain based feature selection method exhibited an accuracy of 87.54% in 350 populations at 100 iterations.

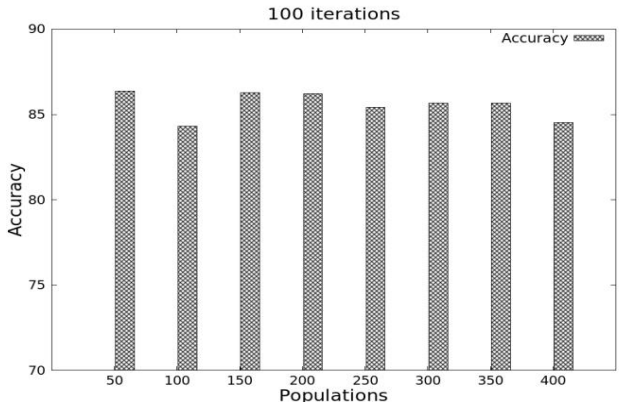

Fig 11: %Accuracy for GA maximizing IG (25 features)

H. Improved Genetic Algorithm for Intrusion Detection System

In this approach authors considered IDS as data analysis process. Here they extract most relevant and effective features 15 features on the basis of information gain in order to reduce the training time and complexity. Because of the attribute subset reduction this system is feasible to apply in real time manner. After the feature reduction they fuzzily input the feature using the triangular function and then apply to the genetic algorithm. The main aim of this approach is to generate rules to differentiate between attack class and normal class type. [8]

In the experimental result they consider two experiment, in first they take threshold value 0.5 and performance measure on different dataset and in second experiment they take threshold value 0.6 and measure the performance.

Type	Training dataset		Testing dataset	
	Detection rate	False positive	Detection rate	False positive
Normal	96.86	3.1	91.78	9.21
attack	97.46	2.5	91.88	18.06

Table 2: Result on different data sets when T=0.5

Type	Training dataset		Testing dataset	
	Detection rate	False positive	Detection rate	False positive
Normal	95.3	4.7	90.51	9.21
Attack	95.05	4.9	81.1	18.06

Table 3: Result on different data sets when T=0.6

From result of second experiment, it is clear that the detection rate is little lower than the accuracy obtained in experiment 1. The detection rates could be higher if the fitness function and the GA parameters were chosen more appropriately.

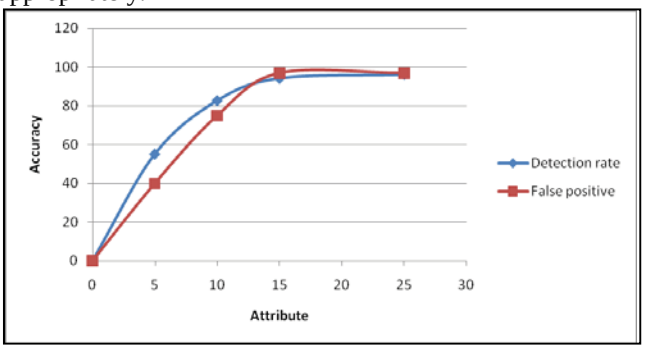

Fig. 12: Attribute vs. Accuracy graph

In this graph which is represent the variation of accuracy with respect to attribute. In this graph accuracy increases gradually with number of attribute but after attaining a certain level the increase in accuracy is very low, on the flip side number of attribute rise sharply. Therefore here little improvement in accuracy is given.

System/ Protocol	Proposed Method/ Idea	Pros	Cons
Using Genetic Algorithm Approach in Intrusion Detection Systems Techniques: survey	IDS systems classification genetic algorithm and their parameters, and different work done on genetic approach in IDS.	---	---
survey On Intrusion Detection Methods	various IDS methods and fuzzy clustering for IDS.	effective for outlier detection	---
IGIDS: Intelligent Intrusion Detection System Using Genetic Algorithm	GA for pruning best individuals in the rule set database ,search space of the resulting rule set is much compact when compared to the original data set	IDS faster and intelligent. high detection rate with low false positives.	Have to be trained for every new attack, lot of legitimate traffic classified as attack.
Implementing Rule based Genetic Algorithm as a Solution for Intrusion Detection System	GA generate a set of rules that can be applicable to the IDS to identify and classify different types of attack connections.	high-quality accuracy, detect complex attack	Complex, more training time
Effect of Change in Rate of Genetic Algorithm Operator on Composition of Signatures for Misuse Intrusion Detection System	set of classification rules which are generated from a predefined intrusion behavior. apply crossover operation make mutation rate is constant	Crossover too much, good segments of individuals get split. some individuals with high fitness' s to be copied directly to the next population.	lot of mutation breaks good genes and stop them from being passed on
Intrusion Detection System Using Genetic Algorithm	GA approach with an improved initial population and selection operator, to efficiently detect various types of network intrusions	increases the performance of the detection rate reduces the false positive rate	DARPA data set need to enrich for detection
Genetic algorithm with Different Feature Selection Method for Intrusion Detection	GA for discovering the most dominant features for classification. IDS with various feature selection methods like information gain, mutual correlation, and cardinality of features.	Information gain based feature selection method exhibited an accuracy of 87.54%	---
Improved Genetic Algorithm for Intrusion Detection System	provides an intrusion detection system (IDS), by modifying the genetic algorithm to network intrusion detection system applied attribute subset reduction on the basis of Information gain.	training time and complexity reduced, Generated rule can detect attack with more efficiency	generated rules were biased to the training dataset

Table 4: Comparison of different IDS using Genetic Algorithm

IV. Conclusion

From this all survey and basic information intrusion detection system is an essential and important technology in Network Security. IDS system is detect intruder which is harmful to the system. Here from all survey we can conclude that using genetic algorithm for IDS we can increase detection rate and reduce false positive rate. Genetic algorithm also use for the feature extraction so using that we can reduce training time also.

In future we can enhance the performance of the system and increase the security of the system also predict the intruders and provide higher security.

References

[1] Owais, Suhail, et al. "Survey: using genetic algorithm approach in intrusion detection systems techniques." Computer Information Systems and Industrial Management Applications, 2008. CISIM'08. 7th. IEEE, 2008.

[2] Mallissery, Sanoop, Jeewan Prabhu, and Raghavendra Ganiga. "Survey on intrusiondetection methods." Advances in Recent Technologies in Communication and Computing(ARTCom 2011), 3rd International Conference on. IET, 2011.

[3] Srinivasa, K. G., et al. "IGIDS: Intelligent intrusion detection system using genetic algorithms." Information and Communication Technologies (WICT), 2011 World Congress on. IEEE, 2011.

[4] Akbar, Shaik, K. Nageswara Rao, and J. A. Chandulal. "Implementing rule based genetic algorithm as a solution for intrusion detection system." Int. J. Comput. Sci. Netw. Secur 11.8 (2011): 138.

[5] Goyal, Mayank Kumar, Alok Aggarwal, and Neelam Jain. "Effect of change in rate of genetic algorithm operator on composition of signatures for misuse intrusion detection system." Parallel Distributed and

Grid Computing (PDGC), 2012 2nd IEEE International Conference on. IEEE, 2012.
[6] Benaicha, Salah Eddine, et al. "Intrusion detection system using genetic algorithm." Science and Information Conference (SAI), 2014. IEEE, 2014.
[7] Cleetus, Nimmy, and K. A. Dhanya. "Genetic algorithm with different feature selection method for intrusion detection." Computational Systems and Communications (ICCSC), 2014 First International Conference on. IEEE, 2014.
[8] Pal, Dheeraj, and Amrita Parashar. "Improved Genetic Algorithm for Intrusion Detection System." Computational Intelligence and Communication Networks (CICN), 2014 International Conference on. IEEE, 2014.

