Review: Soft Computing Techniques (Data-Mining) On Intrusion Detection

[1,]Shilpa Batra , [2,]Pankaj Kumar, [3,]Sapna Sinha
[1,2,3,] *Amity School of Engineering and Technology, Amity University, Noida, U.P.*

Abstract:
With the tremendous growth of various web applications and network based services, network security has become an alarming issue in the vicinity of IT engineers. As the numerous amazing services come to the clients, so does the extensive growth of hackers on the backend. Intrusion poses a serious security risk in the networking environment. Too often, intrusion arises havoc in LANs and heavy loss of time and cost of repairing them. It is said that "prevention is better than cure", so intrusion prevention systems (IPS) and the intrusion detection system (IDS) are used. In addition to the well established intrusion prevention schemes like encryption, client authorization and authentication, IDS can be viewed as a safety belt or fence for network framework. As, the use of interconnected networks have become common, so to have world-wide reports of vulnerabilities and intrusive attacks on systems have increased. CERT noted that between 2000 and 2006 over 26,000 distinct vulnerabilities were reported. An intrusion, which is the set of actions that compromise the integrity, confidentiality, or availability of any resource, and generally exploits one or more bottlenecks of the network. In this paper, we describe various data mining approaches applied on IDS that can be used to handle various network attacks and their comparative analysis.

Keywords: *IDS (intrusion detection system), DOS (denial of service), U2R (user to root), R2L (remote to local), ANN (artificial neural network), MLP (multi-layer perception), GA (genetic*

I. INTRODUCTION

In the recent study, it is being explored that the size of the data has increased worldwide. The need for the larger and larger database has increased. Numerous social networking sites namely Facebook, Twitter etc. use server and databases which stores information of users that are confidential. The latest trend has evoked the use of net banking in which highly confidential data of user is transacted[2]. As the confidential data has grown on network, network security has become more vital. Despite of using prevention techniques like firewalls and secure architecture screening, IDS plays a pivotal role. IDS acts as a burglar or theft alarm which rings whenever a thief tries to steal or hack any data over network. It is used to inform the SSO (Site Security Officer) to defend and take appropriate action in response to the attack; .IDS strengthens the perimeter of the laid network.There are various kinds of attackers but generally can be classified as two:- one who tries to hack the password and steal the user information and the other who exploits its privileges at user level and want to play with the system resources like files, directories and configuration etc.

Network attacks could be:-

[1] DOS (Denial of service):- It aims at limiting the server to provide a particular service to its clients by flooding approach. The general method to do this is ping of death, SYN flags, overloading the target machine.

[2] Probing: - It aims to achieve the computer configuration over network. This can be done by port scanning and port sweeping.

[3] User to Root (U2R) attack: - Its motto is to access the administrator or super user privileges from a given user on which the attacker has previously been given the user level access.

[4] Remote to Local (R2L) attack: - It aims to access the machine by a user which he cannot access by sending message packets to it in order to expose its vulnerabilities. Example guest passwords send mail etc.ADAM (AUDIT DATA ANALYSIS AND MINING) is an intrusion detection built to detect attacks on network level. It uses training and actions mechanism. It classifies the cluster of attacks and alarms on abnormal behaviour of the network.

There are basically two major principles of intrusion detection namely anomaly detection and signature based detection. The former method focuses on analysing the unexpected or abnormal behaviour of the system attributes like deviated CPU cycles, abnormal output to a service requested by a client etc. The issue with using this approach is that, it is difficult to compute the granular attribute characteristics; which are actually time consuming and high rate of false alarm as it is difficult to built tools for such typical and to the depth analysis[6].The latter focuses on determining some predefined signatures or footprints of the attackers who have previously attempt to hack the system. These signatures are stored in KDD (knowledge databases). It is used by expert systems to analyse the previously experienced attacks; but the problem is that it is difficult to keep the records of the type of attacks, the attackers and the issue of storing and maintaining such a huge database of footprints up to date.

Now, we come to the issue of how to implement IDS[5], .the pivotal issue is what to observe while detecting the intrusion and that the source that gives rise to the attack. For this purpose, we take traces to analyse through network log files or also called audits.Moving on to the nature of the source or stimulus of attack, we come to how to observe the stimulus. For observing the point of view, we will use the security log audits, but it is actually a cumbersome and frustrating task. As understanding logs that we need to observe are not actually getting all the necessary traffic that we need, but a flood of network traffic that might be unusable to us. Furthermore, it is not yet concluded that what type of traffic is useful in what kind of circumstances[8][10].This brings us to the results of security logging-what can we observe or what we suspect to observe. Precisely committing the log files is the aim. Henceforth, we need to govern or built rules to classify the data packets in the routes to make IDS successful and effective. As the detector is used to sense the attack occurring in IDS, we need to make an effective decision making IDS. The detector uses various approaches to react to an attack and in the light of this, the main motivation for taking in depth approach to different kinds of detectors that has been deployed on different networking environments.In this paper, we will cover various data mining approaches that underlie the detector principles and mechanisms to react to the different kind of attacks and network layouts.

II. SYSTEM ATTRIBUTES AFFECTING IDS

These are the features that do not affect the detecting principles directly[13]. This divides the cluster of systems based on their approaches to detecting the intrusion in audit data. Following are the vital points:-

[1] Detection time: - It covers two main genres, first: those attackers who try to attack in real time and these need online data analysis and mining them. Second: It processes data with some delay that is non-real time or offline. Although the online analysis can be time delaying to some extent but its computation is much faster than offline.

[2] Stimulus of attack: - The source of attack is considered here. The data for analysis can be taken form two resources: network logs and host logs. The network logs are implemented in NIDS (Network intrusion detection systems) and host logs are used in HIDS (host based intrusion detection).The host log contains kernel logs, application program logs etc. The network log contains the filtered traffic form equipments like routers and firewalls.

[3] Depth of data processing: - The mechanism of data processing could be either continuous data processing or batch processing. In continuous processing all the data traffic running is processed together. While in batch processing, the data is taken in lumps to process. But these terms could be used interchangeably in real time or online data analysis in IDS.

[4] Reaction to the detected attack: - There are two main types of responses to the detected attacks by IDS namely passive and active reactions or response. The passive system responds by notifying on the attack and do not come to remove the affected area of intrusion directly. While the active system comes to eradicate the effect to attack and can be further classified into two categories: firstly, that modifies the state of the attacked system in order to fight back example: terminating the network sessions. Secondly, in response to a detected attack, it attacks back the hacker in order to remove him from his platform.

III. A CLOSER VIEW TO THE DATA MINING APPROACHES IN IDS

3.1 Fuzzy Logic:

It is a form of many valued logic or probabilistic logic that deals with reasoning that is either in true or false form. They range in the degree of 0 or 1. Fuzzy logic is applicable to fuzzy set theory which defines operator on fuzzy set. IF-THEN rules are constructed

The syntax is: IF variable IS event THEN respond The AND, OR & NOT are the Boolean logical operations used. When combined with minimum maximum and compliment, they are called Zadeh operators. Fuzzy relations are stored in the form of relational database. The first fuzzy relation was shown in Maria Zemankoras dissertation. By combining fuzzy logic with data mining the problem of sharp boundary and false positive errors is overcome. This approach can be used with both anomaly as well as signature based IDS. It can be implemented in real time environment. Classification of parameters like SYN flags, FIN flags and RST Flags in TCP headers can be done using fuzzy logic. AN intelligent intrusion detection model integrates fuzzy logic with data mining in two ways; that is fuzzy association rule and fuzzy frequency episode. It integrates both the network level and machine level information. The fuzzy logic represents the commonly found patterns and trends in association rules[1]. For instance occurrence of event X in Y. A variable S (support) tells how often X comes in Y and C (confidence) tells how often Y is associated with X. For example, say sample fuzzy is:

{RN=LOW, SN=LOW}-→{FN=LOW}

C=0.67 & S=0.45, this can be interpreted as SN, FN & RN occurred in 45% of the training sets and the probability of FN occurring at the same time as SN and RN is 67%. In order to implement data mining in anomaly detection approach, mine a set of fuzzy association rules from data set with no anomalies, then given a new data, mine fuzzy association rules on this and compare the similarities of the set of rules mined from new data and normal data.Given a fuzzy episode R: $\{e_1, e_2 \ldots e_{k-1}\} \rightarrow \{e_k\}$, C, S, w; If $\{e_1, e_2 \ldots e_{k-1}\}$ has occurred in the given sequence, then $\{e_k\}$ could be predicted as next to occur event. If the next event does not match any prediction from the rule set then the IDS will alarm it as anomaly. Percentage of anomaly detected can be calculated by the number of anomalies and the number of events occurring. It can be written as:Percentage anomaly= number of anomalies/ number of events.Features selected for IP spoofing and port scanning attacks can be source IP FYN, data size and port number and source IP destination IP, source port and data size respectively.

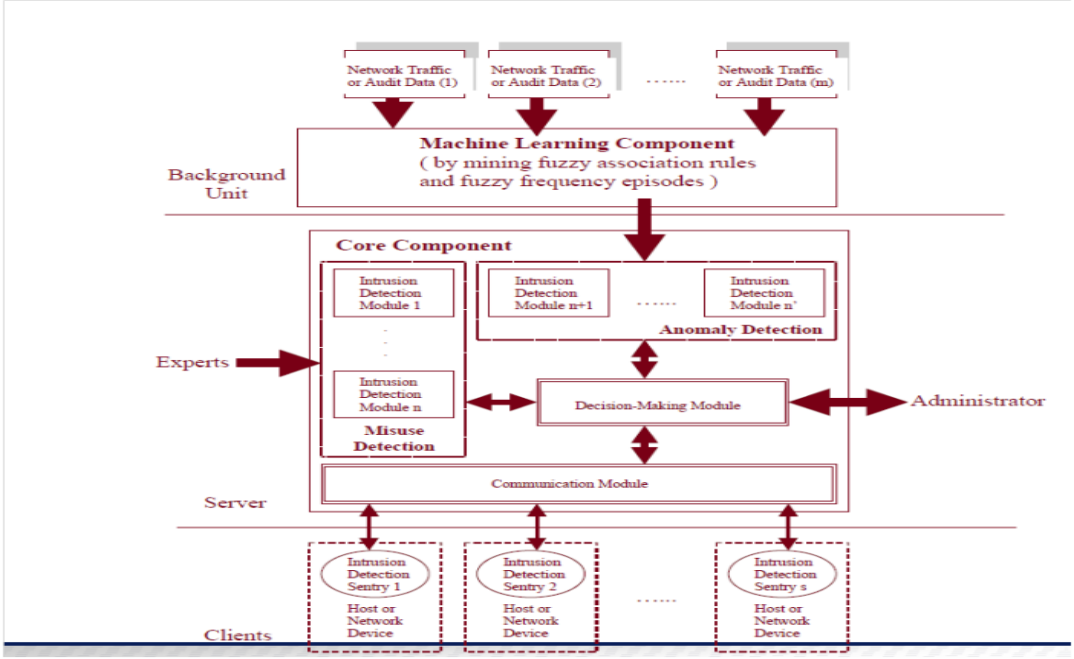

Fig: 1 depicts the implementation of fuzzy rule set and episodes over IDS[1]

3.2 Graph based approach (GrA) :

The graph based approach was developed at UC Davis computer lab that collects data about host and network traffic. Then it aggregates information into activity graphs; that deals with the causal relationship of network activities. The analysis could be done using dedicated hardware like RMON/RMONII. This machine is quick enough to cope up with the speed of network.For the implementation of the graphical data mining approach, we need a supervised network in which modules like packet sniffing, traffic matrix construction, graph clustering, event generation and visualisation are set up. In the graph, the computers in the network layout are represented by node and communication among them is represented by subsequent edges weighted by the amount of data exchanged. Various graph clustering algorithm are then implemented.

Assume we have a graph G with V vertices and E edges then[3]:

G= {V, E} having C_n clusters can be considered as

C1= (V1, V2, V3)
C2= (V4, V5)
:
:
:
Cn = (V_{n-1}, V_n)
Then, G=C1 U C2 U C3 U C4……..Cn
Such that Ci ∩ Cj= Φ

There are various algorithms applied for graph computation namely: hierarchal and non-hierarchal algorithms. The former creates the hierarchy of clusters by subdivision of clusters or combining them.In an agglomerative algorithm, the entire graph forms initially a single cluster which is then subdivided. The latter divides the graph into clusters within one step.Graph visualisation of network traffic is a vital task for planning and managing large network. The network layout is done generally in a geographical way. We cluster the nodes in a structure and this helps to analyse the modification of the network behaviour. A special benefit of visualisation is it is capable of diagnosing modifications in the network structure by building traffic matrices. Changes in network topology, network devices are the reasons for the modifications. Font and colour of nodes indicate changes in their membership to different clusters. It is easy to detect the changes in computer behaviour and information on nodes. The problem with IDS is the rate of false positive and false negative alarms. Therefore, visualization helps in discovering false positive efficiently and reducing the number of false negative. Graph drawing is the task of drawing a given graph on the platform. The tool first places the clusters on the plane. The clusters added form a new graph; this visualisation helps the security manager to build his own opinion on messages from event generation. For event generation, our system collects online network traffic and implements clustering algorithms. Attributes like number of communicating nodes, found clusters ,minima, maxima, out degree and in degree of graph; sink and source node in a cluster; the internal nodes in a cluster and external edges in a cluster are collected. Modifications in a graph could be due to addition of new nodes, lost nodes, splitting of clusters and merging of clusters.CLIQUE and PROCLUS are the methods applied for dimension growth sub-space clustering and dimension reduction, sub space method respectively.

3.3 Neural Networks :

With the rapid expansion of computer networks security has become a very critical issue for computer networks. Various soft computing based methods are being implemented for the development of IDS. A multi-layer training technique is used to evolve a new data domain and offline analysis. Different neural network structure are analysed with regards to the hidden layers. Soft computing is a general term used in context of uncertainty and includes fuzzy logic, AI, Neural networks and genetic algorithms. The idea behind this is to evolve a new and hidden connection records and generalise them. The neural network approach is appropriate for offline data analysis. The training procedure of neural networks is done using validation methods.A non-real time IDS is implemented using multilayer perception (MLP) model. ANN (Artificial Neural Network) is based upon human nervous system processing. It comprises of large number of inter connected processing units called neurons co-ordinating with each other to solve a particular problem. Each processing unit acts upon an activation function. The output of each subsequent layer acts as an input to the next layer. The mechanism of working in ANN is; feed the uppermost layer with input domain and check how closely the actual output for a specific input matches the desired output. Change the weights attached to each layer accordingly. If an unknown input is given to the ANN it presents the output as irrelevant that time but corresponds to that input set.

In IDS, the ANN is implemented by training neurons with the sequence of log audit files and sequence of commands. For the first time when the ANN is fed with current commands it is compared with past W commands (W is the size of window command under examination). Once the ANN is trained with user profile and put into action, it can discover the user behaviour deviation. The next time it is logged in. It is suitable to analyse a small network of computers ranging from 10-15 and analyse a single user command for the whole day. There are numerous commands which describe the user behaviour. Neural networks in the past study were implemented in the UNIX lab for detecting attack specific keywords for host based attacks.A neural network produces two kinds of output in multilayer perception namely normal & abnormal. The output generated is in the form of binary digits 0&1. However, neural network is not capable of identifying attack type. During training phase the neural networks are fed and forward with inputs that are class of network connections and audit logs. The neural network accepts the input processes it through its layered architecture and tries to output the corresponding result. If the output is deviated the neural network gains the knowledge of abnormality and alarms about the attack. In a recent study, data sets contain each event combined with 41 features which were grouped as connection sets, properties of connections etc.

For example, cluster 1 contains the commands used in the connections like file creation, number of root access; cluster 2 includes connection specifications like protocol type service type, duration, number of bytes etc. During investigation it turned out that features like urgent, number of failed logins, is_host_login etc. where playing no role in ID. However, making the data set time consuming and complicated. Therefore, these were removed later. The different possible values were allotted to the rest of the features like TCP=1, UDP=2 etc. The ranges of attributes were different and incompatible and therefore there values were normalised by binary mapping. ANN is efficient to solve a multiclass problem. A binary set approach is used to denote the attack type to feed the neural networks. For example, if an attack is given a value of [0 1 1] and the output generated is [1 1 0]. It is considered as irrelevant. A three layer Neural Network means it has two hidden layers. The uppermost layer is considered as input or buffer layer because no processing task takes place. The cost of neural networks increases as the number of hidden layer increases. But by increasing the number of layers the efficient approximation and accuracy of anomaly detection increases. If we use the neural network of two layers the training cost and time are less. Various tools are used in MATLAB that allows the user to specify the number of layers and activation functions to the layers of ANN[4].

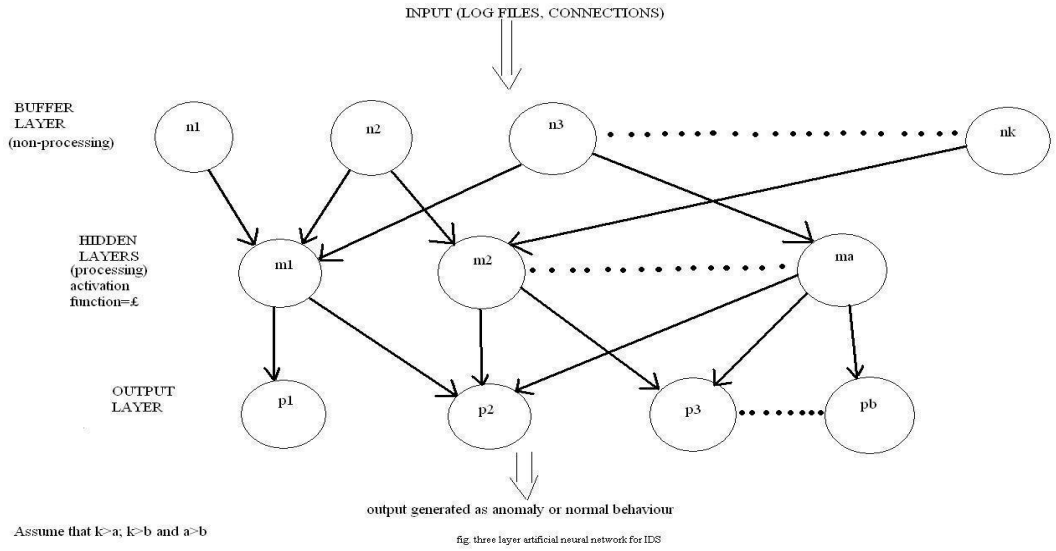

Fig: 2. Working of Neural Network

3.4 Genetic algorithm (GA):

Genetic algorithm is one of the soft computing skills based on the mechanism of evolution and natural selection. The input data set is a set of chromosomes and evolves the next generation of chromosomes using selection, crossover and mutation. The input data set is randomly selected. The problem to be solved is divided into desired input domain. The chromosomes selected are converted into bits, characters or numbers. They are positioned as genes. The set of chromosomes are considered as population. An evaluation function is used to calculate fitness or goodness of each chromosome[7]. Two basic operators are used for reproduction that is crossover and mutation. The best individual chromosome is finally selected for optimization criteria.

The rules used in GA are represented by:

IF {condition} THEN {action};

The conditions to detect the intrusion is generally the current network traffic or connection details like source IP address, destination IP address, port numbers (like TCP, UDP), duration of the connection, protocols used. The action taken in accordance to the security policies are followed by the organisation like alerting the admin by alarm or terminating a connection.

Example IF (the connection having the properties)

Source IP=125.168.90.01;

Destination IP=145.165.10.90;

Destination port=21;

Time=0.2;

THEN (alert by alarming)

Fig.3. Attribute of Genetic Algorithm

The parameters in GA are the evaluation functions which determine whether the connection matches the pre-defined data set and multiply the weights of the field. The matched value ranges from 0 to 1.

Outcome=summation (matched values*weights attached);
Destination IP address is the target of the attack while the source IP is the originator (stimulus) of intrusion. The destination port number indicates the application of the target system to be attacked like FTP, DNS etc.
The suspicious level is the threshold that indicates the extent to which two network connections are considered "matched".
Ω= (outcome-suspicious level);
If a mismatch occurs, the penalty value is computed. The ranking in the equation determines the level of ease of identifying an intrusion represented as:
Penalty= (Ω * ranking / 100);
The fitness of a chromosome is computed as:
Fitness=1 – penalty
The Fitness value ranges from 0 to 1.
The mechanism for GA can be followed as:-
Pass 1: Gather the input set and initialise the population in any order (arbitrarily)
Pass 2: Do the summation of the records
Pass 3: the new population of chromosomes are produced.
Pass 4: Calculate the result by implementing Crossover operator to the Chromosome
Pass 5: Apply Mutation operator to the chromosome
Pass 6: Evaluate Fitness f(x) = f(x) / f (sum)
Where, f(x) is the fitness of individual x
and f is the sum of fitness of all individuals in a pop
Pass 7: Rank Selection Ps (i) = r (i) / rsum
Where, Ps (i) is probability of selection Individual r(i) is rank of Individuals rsum is sum of all fitness values.
Pass 8: Choose the top best 60% of Chromosomes Into new population
Pass 9: if the number of generations is not reached, go to Pass 3[7][11][12].

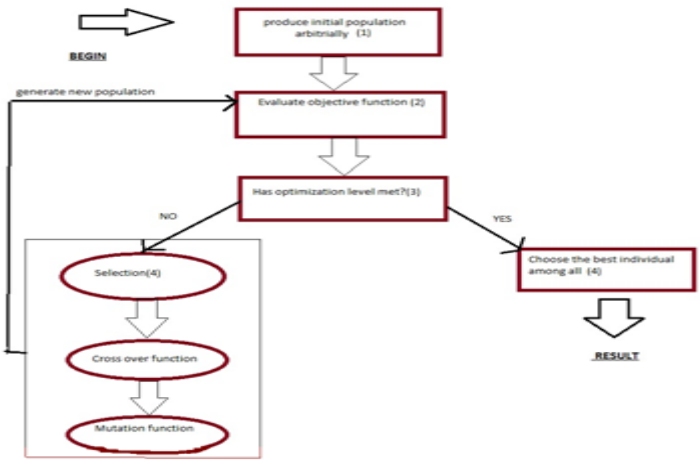

Fig: 4 Working of Genetic Algorithm

As there are various types of attacks which can be listed as:

- DOS attack: Smurf, Neptune, Pod
- U2R:buffer overflow, Perl
- R2L:guest password, ftp write, phf, spy
- Probe: satan, IPsweep, portsweep

There are various samples Rule sets used to determine the types of attacks by using GA approach. For instance:

[1] IF (duration=0 and protocol=ICMP) THEN (smurf)
[2] IF (duration=0 and protocol=TCP and host_srv_count is greater than 1 and less than 128) THEN (Perl (u2R attack))
[3] IF (duration=0 to 289, protocol=UDP and src_bytes=0) THEN (guess password)

There is numerous work done related to GA like: Lu developed a method to determine a set of rule classification with the help of past data of networks; Xiao detected abnormal behaviour of networks by using mutual information and complexity reduction; Li implemented GA by using quantitative features.

IV. CONCLUSION

Fuzzy logic is one of the soft computing technique efficient in implementation of rule based data mining on intrusion over non fuzzy data sets. Although much of the success has been achieved by using this technique yet it is needed to be applied on high speed workstations and misuse rule base is still in progress by using fuzzy association rules. The fuzzy is now days optimised by the integration of genetic algorithm with it[1]. In graph based approach, The visualization helps the security manager to get insight to the current usage of the computer network[3]. He has the possibility to learn more about the reasons for events or warnings from his intrusion detection system. This form of presentation helps detecting false positives. The event generating system has to be improved. The concepts are interesting, but additional work is needed to optimize the process. It is possible to automatically detect anomalies in the communication structure of a surveyed network, but the goal of detecting a large number of different attacks is not yet reached. Integration in the existing intrusion detection system is planned for the near future. Additional graph algorithms, especially clustering algorithms will be tested and compared with the used ones. More features shall be extracted from the clustered traffic graphs and different learning methods will be tested. The visualization will be optimized to the needs of the permanent usage of the system. An approach for a neural network based intrusion detection system, intended to classify the normal and attack patterns and the type of attacks. It should be mentioned that the long training time of the neural network was mostly due to the huge number of training vectors of computation facilities. However, when the neural network parameters were determined by training, classification of a single record was done in a negligible time[4].

Therefore, the neural network based IDS can operate as an online classifier for the attack types that it has been trained for. The only factor that makes the neural network off-line is the time used for gathering information necessary to compute the features. The basic problem with ANN is the over fitting, it is alright to use a small data set but becomes cumbersome as the size of the data increases. The paper presents the Genetic Algorithm for the Intrusion detection system for detecting DoS, R2L, U2R, Probe. The time to get thorough with the features to describe the data will be reduced with a combination of Genetic Algorithm based IDSs. This provides a high rate of the rule set for detecting different types of attacks. The results of the experiments are good with an 83.65% of average success rate and got satisfied. Presently, systems are more flexible for usage in different application areas with proper attack taxonomy. As the intrusions are becoming complex and alter rapidly an IDS should be capable to compete with the thread space. Genetic Algorithm detects the intrusion while correlation techniques identify the features of the network connections. Optimizing the parameters present in the algorithm reduces the training time. More reduction techniques may be referred to get valuable features in future

REFERENCES

[1] Susan M. Bridges, Rayford B. Vaughn 'FUZZY DATA MINING AND GENETIC ALGORITHMS APPLIED TO INTRUSION DETECTION' 23rd National Information Systems Security Conference October 16-19, 2000.
[2] Bertrand Portier, Froment-Curtil Data Mining Techniques for Intrusion Detection.
[3] Jens Tölle, Oliver Niggemann 'Supporting Intrusion Detection by Graph Clustering and Graph Drawing'.
[4] Mehdi Moradi, Mohammad Zulkernine 'A Neural Network Based System for Intrusion Detection and Classification of Attacks' Natural Sciences and Engineering Research Council of Canada (NSERC).
[5] Jungwon Kim, Peter J. Bentley, UWE Aickckelin, Julie Greensmith, Gianni Tedesco, Jamie Twycross 'Immune system approaches to intrusion detection –a review' Natural Computing (2007) 6:413–466 _ Springer 2007DOI 10.1007/s11047-006-9026-4, Springer 200

[6] Wenkee Lee, Salvatore J. Stolfo, Kui W. Mok 'Adaptive Intrusion Detection: A Data Mining Approach' Artificial Intelligence Review 14: 533–567, 2000. issues on the Application of Data Mining. © 2001 Kluwer Academic Publishers. Printed in the Netherlands.
[7] Wei Li 'Using Genetic Algorithm for Network Intrusion Detection'.
[8] Justin Lee, Stuart Moskovics, Lucas Silacci 'A Survey of Intrusion Detection Analysis Methods' CSE 221 spring 1999.
[9] Mikhail Gordeev 'Intrusion Detection Techniques and Approaches'.
[10] Wenke Lee, Salvatore J. Stolfo 'Data Mining Approaches for Intrusion Detection' 7th USENIX Security Symposium, 1998
[11] A.A. Ojugo, A.O. Eboka, O.E. Okonta, R.E Yoro (Mrs), F.O. Aghware 'Genetic Algorithm Rule-Based Intrusion Detection System (GAIDS)' Journal of Emerging Trends in Computing and Information Sciences VOL. 3, NO. 8 Aug, 2012
[12] Detection System (GAIDS)" Journal of Emerging Trends in Computing and Information Sciences ©2009-2012 CIS Journal.
[13] B. Uppalaiah, K. Anand, B. Narsimha, S. Swaraj, T. Bharat 'Genetic Algorithm Approach to Intrusion Detection System' IJCST Vol. 3, Issue 1, Jan. - March 2012 .
[14] Stefan Axelsson 'Intrusion Detection Systems:-A Survey and Taxonomy'14 March 2000