ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 3, Issue 1, January 2014*

# Fuzzy Data Mining Based Intrusion Detection System Using Genetic Algorithm

**Harshna[1], Navneet Kaur[2]**

M.Tech (C.S.E), Department Of Computer Science & Engineering of RIMT Institutions, MandiGobindgarh, Sirhind[1]

Assistant Professor, Department of Computer Science & Eng., of RIMT Institutions, MandiGobindgarh, Sirhind[2]

**Abstract**: Today it is very essential to preserve a high level security to ensure safe and trusted communication of information between various organizations. But due to various threats like intrusions and misuses, secured data communication over internet and any other network is very difficult to achieve. An intrusion can be defined as any set of actions that compromise the three main aims of security i.e integrity, confidentiality or availability of a network resource(such as user accounts, file system, kernels & so on). Data mining plays a outstanding role in data analysis. These systems identify attacks and react by generating alerts or by blocking the unwanted data/traffic. So data mining plays an important role in Intrusion Detection System as it relays upon the auditing of data. Due to the use of fuzzy logic, the system can deal with mixed type of attributes and also avoid the sharp boundary problem. Genetic algorithm is used to extract many rules which are required for anomaly detection systems.

**Keywords**: KDD, Data Mining, Intrusion Detection System, Fuzzy Logic, Genetic Algorithm

## I. INTRODUCTION

KDD stands on the Knowledge Discovery in Databases is the process of finding associations and patterns in raw data automatically from large databases and gives the output results. In particular, the KDD process consists of the following steps:

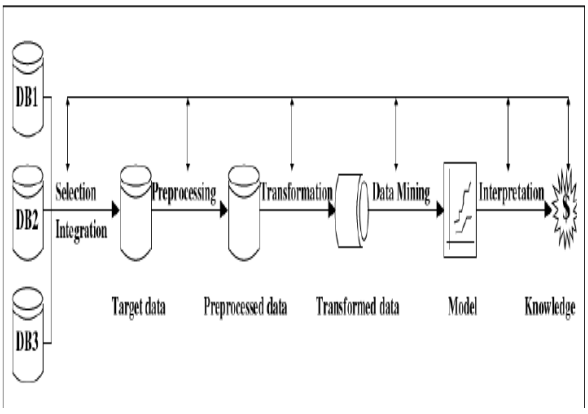



Fig 1: Steps of KDD

1. Selection and Integration: The data that is to be mined may exist in dissimilar and mixed data form. Therefore, It is necessary to first selecting the appropriate data from a variety of databases for analysis and integrate that data values into a logical data store.
2. Pre-processing: Raw data may have invalid or missing values. Invalid values are corrected while misplaced values are supplied or predicted.
3. Transformation: The data are transformed into representations that are suitable for mining tasks.
4. Data Mining: Data Mining is the core part of the KDD process referring to the purpose of intelligent techniques to take out the secreted knowledge from the transformed data.
5. Interpretation and Evaluation: A data mining system has the ability to produce a great number of patterns but only a small fraction of them may be of attention. Therefore the suitable methods are required to estimate the valuable results.

## II. DATA MINING

Data mining is also defined as knowledge discovery in databases(KDD) has attain a great deal of interest in the information industry and in society. Due to the availability of large amount of data and its major need for extracting such data into useful information is increasing rapidly. Various machine learning algorithms, Neural Network, Support Vector Machine, Fuzzy Logic ,Genetic Algorithm and Data Mining have been broadly used to detect intrusive activities both for known and unknown dynamic datasets. Data mining tasks can be classified into 2 categories namely descriptive mining & predictive mining. The descriptive mining techniques like as Association, Clustering , Sequential Pattern discovery, is used to find human interpretable patterns that describe the data. The predictive mining techniques such as classification, Regression, Deviation, detection, etc., are used to predict unknown or upcoming values of other variables.

Specific uses of data mining include:

- Market segmentation - Identify the universal characteristics of customers who buy the similar products from your company.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 3, Issue 1, January 2014*

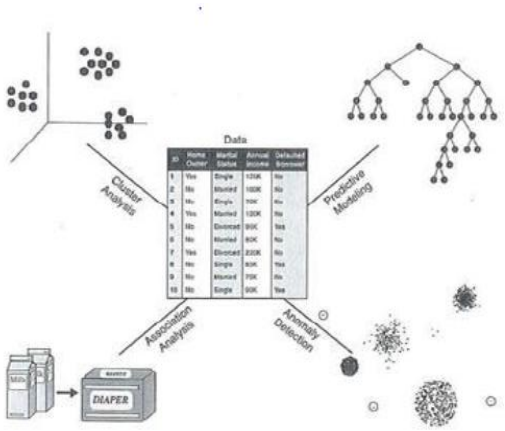



Figure 2: The four core of Data mining task

 Customer churn - calculate which customers are likely to go away from your company and buy the products from your competitor.
 Fraud detection - Identify which transactions are most likely to be fraudulent.
 Direct marketing - Identify which prediction should be included in a mailing list to gain the maximum response rate.
 Interactive marketing - Predict what each individual accessing a Web site is most likely interested in seeing.
 Market basket analysis - Understand and identify the products or services that are usually purchased together; e.g., soap and oil.
 Trend analysis - Reveal the difference between typical customers this month and last

## III. INTRUSION AND DETECTION OVERVIEW

Intrusion is defined as the attempts to bypass the security mechanisms of a computer or network. The basic goals of computer security are integrity, confidentiality, and availability. As integrity involves no duplicity in data, confidentiality means privacy of the data and availability involves the presence of the data in the accurate manner when it is to be required. So, Intrusion is a set of unwanted actions aimed to compromise these security goals. To prevent these actions, intrusion prevention (authentication, encryption, etc.) alone is not sufficient. So before Intrusion prevention , Intrusion detection is needed. The following section give a short overview of networking attacks:

Networking Attacks
Every attack on a network can comfortably be placed into one of the following groupings [1]. The overview of the four major categories of networking attacks are described as below as:

a. Denial of Service (DoS)
A DoS attack is a type of attack in which the hacker makes a computing or memory resources too busy or too full to serve legal networking requests and hence denying users access to a machine. The main examples are neptune, ping, apache, smurf, of death, back, mail bomb, UDP storm etc. are all DoS attacks.

b. Remote to User Attacks (R2L)
A remote to user attack is about an attack in which a user sends packets to a machine over the internet, which she/he does not have access to in order to expose the machines vulnerabilities. It exploits privileges which a local user would have on the computer e.g. xlock, guest, xnsnoop, phf, sendmail dictionary etc.

c. User to Root Attacks (U2R)
These attacks are exploitations in which the hacker starts off on the system with a normal user account and attempts to abuse the legal actions in the system in order to gain super user privileges e.g. perl, xterm.

d. Probing
Probing is defined as an attack in which the hacker scans a machine or a networking device in order to determine weaknesses or vulnerabilities that may later be exploited so as to compromise the system. This technique is commonly used in data mining e.g. saint, portsweep, mscan, nmap etc.

## IV DATA MINING MEETS INTRUSION DETECTION

a. Anomaly Detection or Profile Matching *:* This technique is based on the normal behaviour of a subject (e.g., a user or a system); any action that significantly deviates from the normal behaviour is considered as an intrusive action. Misuse detection catches intrusions in terms of the characteristics of known attacks or system vulnerabilities; any action that conforms to the pattern of a known attack or vulnerability is considered intrusive. The anomaly approach is focused on normal behaviours patterns. When a new kind of activity becomes acceptable (does not contradict to security policy), the normal behaviour pattern database must be updated; otherwise the activity will be treated as an intrusion and will result in false positives. Attacks and deviations from normal activity are anomaly by definition and deserve the IDS user's attention.

Although anomaly detection can find out unknown patterns of attacks, it also suffers from several drawbacks. A general problem of all anomaly detection approaches, with the exception of the specification-based technique, is that the subject's normal behaviour is modelled on the basis of the (audit) data collected over a period of normal operation. If undiscovered intrusive activities occur during this period, they will be taken as normal activities. In addition, because a subject's normal behaviour usually changes over time (for example, a user's behaviour may change when he moves from one project to another), the IDSs that use the above approach usually allow the subject's profile to gradually change. So, this gives an intruder the chance to gradually train the IDS and trick it into accepting intrusive activities as normal. Also,

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 3, Issue 1, January 2014*

because these approaches are all based on summarized information, they are insensitive to stealthy attacks. Because of some technical reasons, the current anomaly detection approaches usually suffer from a high false-alarm rate. Another difficult problem in building such models is how to decide the features to be used as the input of the models (e.g., the statistical models). In the existing models, the input parameters are generally decided by domain experts (e.g., network security experts) in ad hoc ways. So, it is not guaranteed that all the features related to intrusion detection will be selected as input parameters. Missing important intrusion-related features makes it difficult to distinguish attacks from normal activities, having non-intrusion-related features could introduce "noise" into the models and thus affect the detection performance.

4.2 Misuse Detection or Signature Matching: Misuse detection is said to be complementary to anomaly detection .In misuse detection approach, firstly abnormal system behaviour is defined, and then define any other behaviour, as normal behaviour. Its main advantage is simplicity of adding known attacks to the model. Therefore, this systems look for well-defined patterns of known attacks or vulnerabilities. They can catch an intrusive activity even if it is so negligible that the anomaly detection approaches tend to ignore it. Attacks and deviations from normal behaviour are taken as anomalies.

The disadvantage of misuse detection is that it cannot detect novel or unknown attacks. As a result, the computer systems protected solely by misuse detection systems face the risk of being comprised without detecting the attacks. In addition, due to the requirement of explicit representation of attacks, the detection system requires the nature of the attacks to be well understood. It implies that human experts must work on the analysis and representation of attacks. So, it is time consuming and error prone.

Additionally, intrusion detection systems (IDSs) are categorized according to the kind of input information they analyze. This leads to the distinction between host-based and network-based IDSs. Host-based IDSs analyze host-bound audit sources such as operating system audit trails, system logs, or application logs. Network-based IDSs analyze network packets that are captured on a network.

## V DATA MINING AIDS IN INTRUSION DETECTION

Well-known data mining techniques used for intrusion detection are below as:

- Classification
- Clustering
- Association-Rule mining

A. Classification

Classification is one of the important technique of data mining. Its goal is to build the classification attribute model based on attributes of the data. Data classification has two steps.

The first step is inspired from the supervised learning process. In this step, a data set is selected. The class label of each set (training samples) for training data set is known. The class label of each training samples is provided. Usually, the learning model is described by the classification rules, mathematical formula or decision tree. The second step, the model is classified. First the prediction accuracy of the model (classification rules) is evaluated. Then, for each test sample, the known class label and the prediction label of the sample are compared. If the model's exactness rate can be accepted, it will be used to classify the data set that the class label is un-known.[9]

B. Clustering

Clustering is the process to identify the internal rules of the data object. The objects are grouped to form a class of related objects i.e clusters, and export the data distribution. Similar or dissimilar measure is based on the values of the property defined by the data object. Usually, it is defined by the distance. When the mining task is confronted with the lack of domain knowledge or incomplete data set, clustering is used to divide the unknown data object into different classes automatically. Distinction between classification and clustering is that classification is applied to the data object, and clustering is to find the classification rules unstated in the mixed data objects.

C. Association rules

Association rule mining is one of the most well-known approaches in data mining techniques. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. It is to find the exciting connections between items of a given data set. Suppose Database T is a collection of n transactions, {T1, T2, . . ., Tn} and I is the set of all items, {i1, i2, . . ., im}, where each of the transactions Tj($1 \leq j \leq n$) in the database T represents a set of items (Tj $\subseteq$I). An item set is defined as a non-empty subset of I.

An association rule can be represented as: $X \rightarrow Y(c, s)$, where $X \subseteq I$, $Y \subset I$ and $X \cap Y = \varphi$. In this association rule, s is called support and c is confidence of the association rule. The support is the percentage of the transactions in which both X and Y appear in the same transaction and the confidence is the ratio of the number of transactions that contain both X and Y to the number of trans-actions that contain only X. It can be described as follows:

Support $(X \rightarrow Y) = P (X \cup Y)$
Confidence $(X \rightarrow Y) = P (Y|X)$

Association rules were first developed to find correlations in transactions using retail data . For example, if a customer who buys a soft drink (A) usually also buys potato chips (B), then potato chips are associated with soft drinks using the rule A--> *B.* Suppose that 25% of all customers buy both soft drinks and potato chips and that

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 3, Issue 1, January 2014*

50% of the customers who buy soft drinks also buy potato chips. Then the degree of support for the rule is s = 0.25 and the degree of confidence in the rule is c = 0.50.

## VI. FUZZY LOGIC

Fuzzy Logic is derived from fuzzy set theory dealing with reasoning that is approximate rather than precise. Florez G. et al. applied an improved algorithm of the fuzzy data mining approach to the IDS. The fuzzy data mining technique is used to extract the patterns that represent normal behaviour for intrusion detection. Luo J. also attempted classification of the data using Fuzzy logic rules. Typically, an IDS uses Boolean logic in determining whether or not an intrusion is detected and the use of fuzzy logic has been investigated as an substitute to Boolean logic in the design and implementation of these systems. Fuzzy logic focuses on the formal principles of approximate reasoning . It provides a sound foundation to handle the mechanisms using varying degrees of truth. As boundaries are not always clearly defined, fuzzy logic can be used to identify complex pattern or behavior variations. This is done by building an Intrusion Detection System that combines fuzzy logic rules with an expert system in charge of evaluating rule truthfulness. Fuzzy logic is significant for the intrusion detection problem for two major reasons. First, many quantitative features are involved in intrusion detection. Security-related data categorizes the statistical measurements into four types: ordinal, categorical, binary categorical, and linear categorical . Both ordinal and linear categorical measurements are quantitative features that can potentially be viewed as fuzzy variables. Two examples of ordinal measurements are the CPU usage time and the connection duration. An example of a linear categorical measurement is the number of different TCP/UDP services initiated by the same source host. The second motivation for using fuzzy logic to address the intrusion detection problem is that security itself includes fuzziness. Given a quantitative measurement, an interval can be used to denote a normal value. Then, any values falling outside the interval will be considered anomalous to the same degree regardless of their distance to the interval. The same applies to values inside the interval, i.e., all will be viewed as normal to the same degree. The use of fuzziness in representing these quantitative features helps to smooth the abrupt separation of normality and abnormality.

## VII. FUZZY ASSOCIATION RULES

As per the different quantitative attributes, association rule is divided into Boolean association rules and quantitative association rules. In reality, the data are quantitative in most cases, so the quantitative association rules mining research is very important. The general method to solve quantitative association rules is that the value of the property is divided into several regions by a certain criteria and then is converted to a sequence-<attribute, interval>. Thus quantitative association rule will be transformed into Boolean association rules. How-ever, there are some problems. On the one hand, if the interval division is too large, confidence of the rules included in the interval will be very low. So that it will cause a small number of rules, and will be a corresponding reduction in the amount of information. If the interval division is too small, support of the rules included in the interval will be very low. So that it will cause a small number of rules. On other hand, if the domain of property is divided into the non-overlapping interval, the discrete data in the database is mapped to the interval. As potential elements near the interval are excluded by clear division, it will lead to some significant interval is ignored. If the domain of property is divided into overlapping intervals, the elements in the border may be in two intervals at the same time. These elements will contribute to the two intervals, resulting in some intervals are overemphasized. In order to solve the problem of sharp boundary, fuzzy theory is proposed. The membership function is used to define data set in fuzzy sets of the attribute domain, in order to achieve the purpose of softening the border.

### A. Fuzzy Association Rules

So to overcome the drawback i.e sharp boundary problem of association rules, they are integrated with the Fuzzy logic, so become Fuzzy Association rules. Given a database T with attributes I and the definitions of fuzzy sets associated with attributes in I, the objective is to find out some interesting regularities between attribute values in a guided way. We are combining techniques from fuzzy logic and data mining for our anomaly detection system. The advantage of using fuzzy logic is that it allows one to represent concepts that could be considered to be in more than one category (or from another point of view—it allows representation of overlapping categories). In standard set theory, each element is either completely a member of a category or not a member at all. In contrast, fuzzy set theory allows partial membership in sets or categories. The second technique, data mining, is used to automatically learn patterns from large quantities of data. The integration of fuzzy logic with data mining methods helps to create more abstract and flexible patterns for intrusion detection.

## VIII. GENETIC ALGORITHM

A Genetic Algorithm (GA) is a programming technique that mimics biological evolution as a problem-solving strategy . It is based on Darwinian's principle of evolution and survival of fittest to optimize a population of candidate solutions towards a predefined fitness . GA uses an evolution and natural selection that uses a chromosome-like data structure and evolve the chromosomes using selection, recombination and mutation operators . The process usually initiated with randomly generated population of chromosomes, which represent all possible solution of a problem that are considered candidate solutions. From each chromosome different positions are encoded as bits, characters or numbers.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 3, Issue 1, January 2014*

These positions could be referred to as genes. When we use GA for solving various problems three factors will have vital impact on the effectiveness of the algorithm and also of the applications . They are: i) the fitness function; ii) the representation of individuals; and iii) the GA parameters. In intrusion detection, the GA is employed to derive a set of Association rules from network audit data, and the support-confidence framework is utilized as a fitness function to judge the quality of each rule. Good properties of GA are it is robust to noise, self learning capabilities, no gradient information is required to find the global optimal or sub-optimal solution. High attack detection rate and low false-positive rate are the advantages of GA techniques . Therefore we have used G.A for Intrusion Detection.

We implement this in the MATLAB environment. The name MATLAB stands for matrix laboratory. MATLAB is a numerical computing environment and fourth-generation programming language. MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation.

Working steps of Genetic Algorithm are:

1. [START] Generate random population of n chromosomes i.e. suitable for the problem.
2. [FITNESS] Evaluate the fitness f(x) of each chromosome x in the population.
3. [NEW POPULATION] Create a new population by repeating following steps until the new population is complete.

a) [SELECTION]: Reproduction (or selection) is an operator that makes more copies of better strings in a new population. Reproduction is usually the first operator applied on a population .

b) [CROSSOVER]: A crossover operator is used to re-combine two strings/parents to get better new two strings/children. It is important to note that no new strings are formed in the reproduction phase. In the crossover opera-tor, new strings are created by exchanging information among strings of the mating pool.

c) [MUTATION]: Mutation adds new information in a random way to the genetic search process . It is an operator that introduces diversity in the population whenever the population tends to become homogeneous due to repeated use of reproduction and crossover operators .

d) [ACCEPTING] place new offspring in the new population.

4. [REPLACE] use new generated population for the further run of the algorithm.
5. [TEST] if the end condition is satisfied then stops and re-turns the best solution in current population.
6. [LOOP] Go to step 2.

## IX. RESULTS

The results obtained from the fuzzy- G.A based algorithm on the given dataset of a network which is comprised of numeric data. The analytical results of the algorithm under the following parameters are implemented .

a. Elapsed Time: It is defined as the completion time of the algorithm. Lesser the elapsed time, less time to execute the algorithm more efficient is the algorithm. Elapsed time is calculated by measuring the finishing time of the exit task by the algorithm.

b. No. of Iterations: It is defined as the number of loops to be executed to run the whole algorithm.

c. Fitness function: It is defined as a threshold value to select the required nodes.

d. Membership Function: When we use fuzzy logic, a particular value can have a degree of membership between 0 and 1 and can be a member of more than one fuzzy set. So, M.F describes the degree of membership of a particular node in the network.

e. Population size: It is defined as the total number nodes present in the system or network.

f. Correlation Ratio: It is defined as ratio of the a same term or element to be find out in the given iteration.

1. Results after observing the relationship b/w the no.of iterations, Membership func. and Time

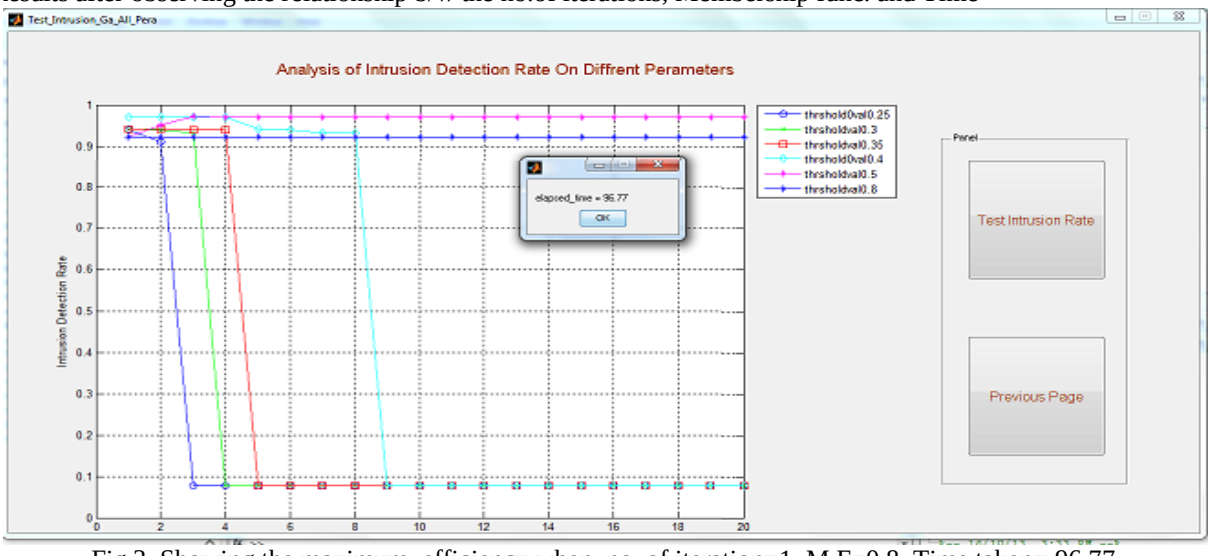
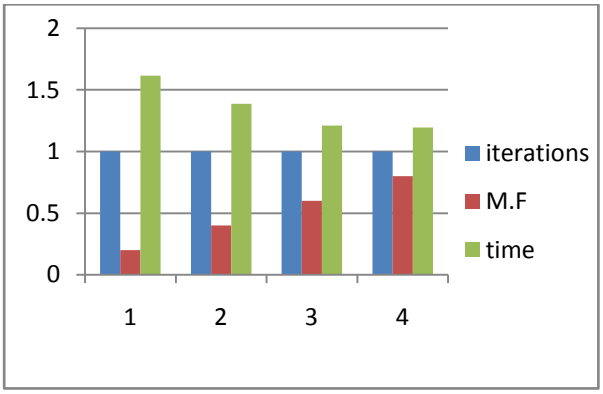



Fig 3. Showing the maximum efficiency when no. of iteration=1, M.F=0.8, Time taken= 96.77

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 3, Issue 1, January 2014*

| Iterations | M.F | Time |
|---|---|---|
| 1 | 0.2 | 1.614833 |
| 1 | 0.4 | 1.386667 |
| 1 | 0.6 | 1.210333 |
| 1 | 0.8 | 1.195 |




2. Results after observing the relation b/w No.of Iterations, population Size and Time

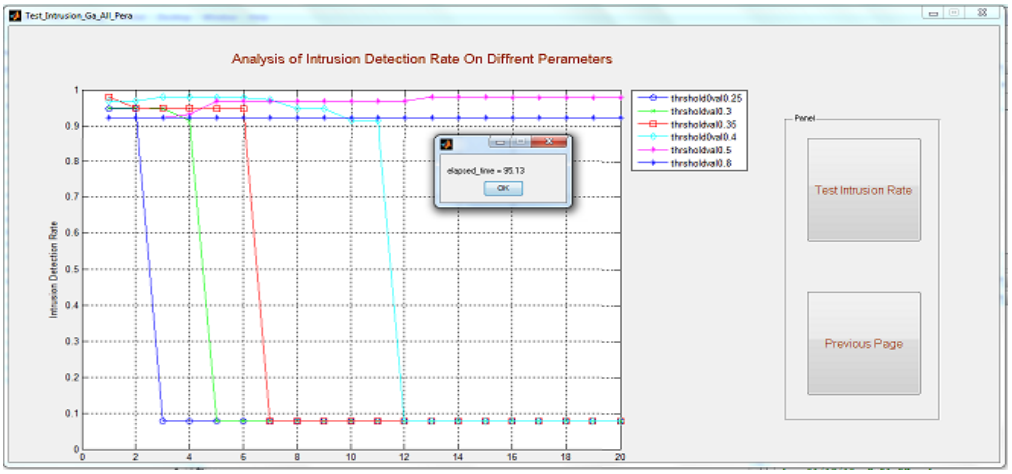
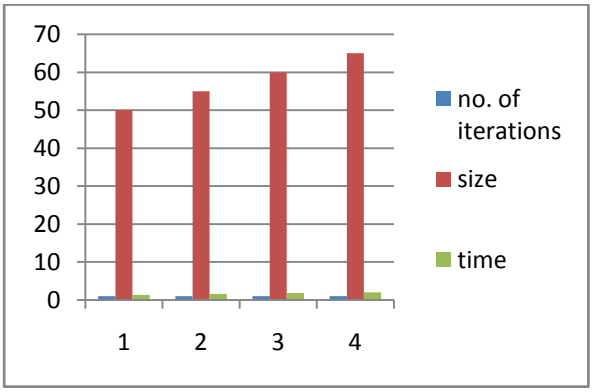



Fig 4: Showing Maximum efficiency when No. of iterations=1, Size = 55 and Time = 1.5805

| Iterations | size | time |
|---|---|---|
| 1 | 50 | 1.371 |
| 1 | 55 | 1.5805 |
| 1 | 60 | 1.826 |
| 1 | 65 | 2.0045 |




3. Results after observing the relationship b/w no. of iterations, correlation ratio and Time

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 3, Issue 1, January 2014*

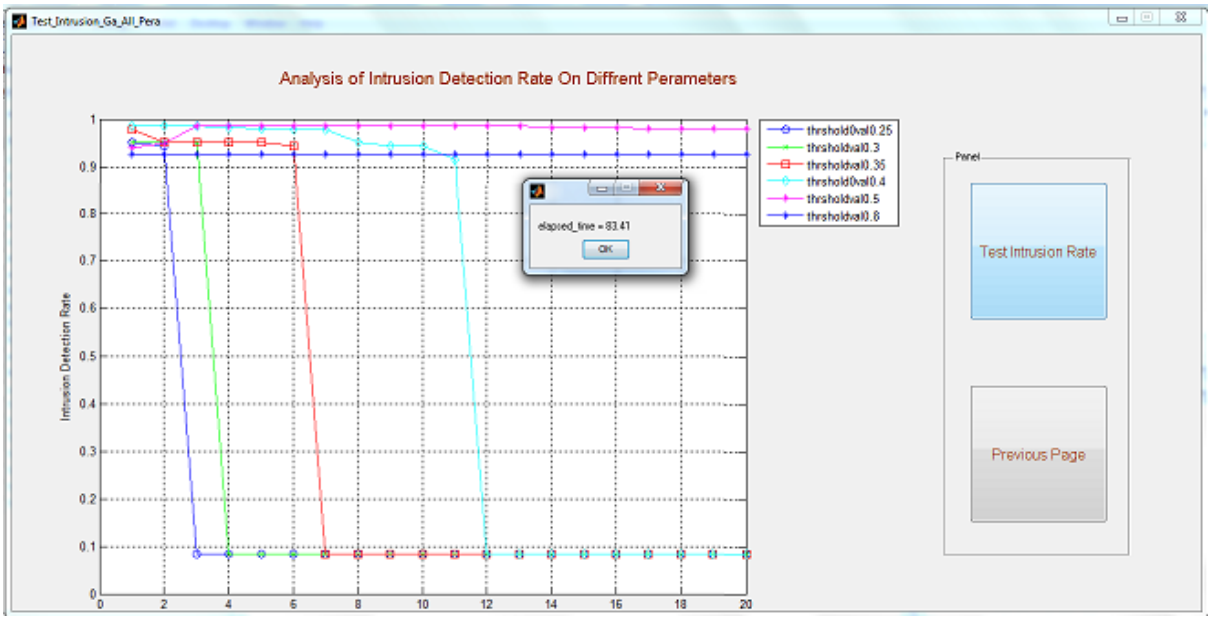
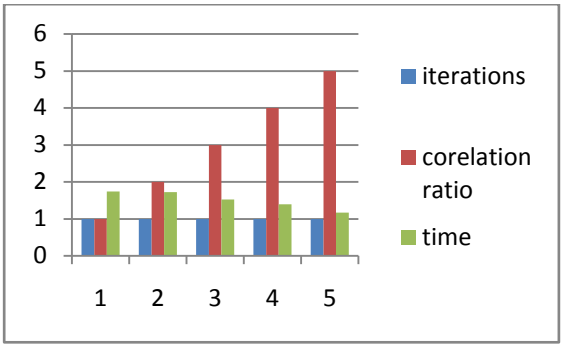



Fig 5: Showing efficiency when no. of iterations = 1, C.R= 4, time= 1.390

| Iterations | Correlation Ratio | Time |
|---|---|---|
| 1 | 1 | 1.742833 |
| 1 | 2 | 1.724 |
| 1 | 3 | 1.523667 |
| 1 | 4 | 1.390167 |
| 1 | 5 | 1.168667 |




## X. CONCLUSION

Intrusion Detection is one of the major issue in any computer networks environment. Various methods related to intrusion detection system are studied .Association Rule Mining is used for intrusion detection in this paper. Use of fuzzy logic overcomes the sharp boundary problem caused by the association rules. Thus fuzzy association rules can be mined to find the abstract correlation among different security features. Using genetic algorithms with the fuzzy data mining method may result in the tune of the fuzzy membership functions to improve the performance and select the set of features available from the audit data that provide the most information to the data mining component. These algorithms are mostly used for optimization problems. Therefore, integration of fuzzy logic with association rules and GA generates more abstract and flexible patterns for intrusion detection that can be used for both misuse and anomaly detection. The input parameters of any algorithms to get the best results are very carefully taken. This algorithmis implemented on the numeric dataset. The execution time depends upon the how large the no.of iterations, population size , correlation ratio. Efficiency increases with the increase in execution time till a certain limit. But one can do work on time management in future to make it more efficient.

## ACKNOWLEDGEMENT

I express my sincere gratitude to my guide Ms Navneet Kaur, for her valuable guidance and advice. Also I would like to thanks all the faculty members and colleagues for their continuous support and encouragement.

## REFERENCES

[1] A. Sung, S. Mukkamala, Identifying important features for intrusion detection using support vector machines and neural networks in Symposium on Applications and the Internet, pp. 209–216. 2003.
[2] Agrawal R. and Srikant R.,Fast algorithms for mining association rules, in Proceeding 20th VLDB Conference, San-tiago, Chile, pp. 487–499, 1994.
[3] Anderson.J.P, Computer Security Threat Monitoring & Surveilance, Technical Report, James P Anderson co., Fort Washington, Pennsylvania, 1980.
[4] Denning D.,An intrusion detection model," IEEE Trans. Software Eng., vol. 13, no. 2, pp. 222–232, Feb. 1987.
[5] Ektefa M., Memar S.,Intrusion Detection Using Data Mining Techniques,IEEE Trans., 2010.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021



[6] Florez G., Bridges S., Vaughn R., An Improved Algorithm for Fuzzy Data Mining for Intrusion Detection, Annual Meeting of The North American Fuzzy Information Processing Society Proceedings, 2002.
[7] Goldberg D.,Genetic Algorithm in Search, Optimization and Machine Learning, Reading, MA: Addison-Wesley, 1989
[8] Jian Pei, Upadhayaya.S.J, Farooq.F, Govindaraju.V,Data Mining for Intrusion Detection: Techniques, Applications & Systems, in the Proceedings of 20th International Conference on Data Engineering, pp-877-887, 2004.
[9] Lee W. and Stolfo S.,Data Mining Approaches for Intrusion Detection,Computer Science Department Columbia University.
[10] Luo J., Integrating fuzzy logic with data mining methods for intrusion detection, Master's Thesis, Department of Computer Science, Mississippi State University, Starkville, MS, 1999.