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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

Fuzzy-Association Rule Mining based Intrusion Detection System using Genetic Algorithm

Harshna[1], Navneet Kaur[2]

M.Tech, Department Of Computer Science & Engineering of RIMT Institutions, MandiGobindgarh, Sirhind[1]

Assistant Professor, Department of Computer Science & Engineering 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. So Intrusion Detection Systems have become a needful component in terms of computer security. 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. So data mining plays an important role in Intrusion Detection System as it relays upon the auditing of data. These systems identify attacks and react by generating alerts or by blocking the unwanted data/traffic. The proposed work includes fuzzy logic with a data mining method which is a association rule mining method based on genetic algorithm. 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. An association-rule- mining method is used to extract a sufficient number of important rules for the user's purpose rather than to extract all the rules meeting the criteria which are useful for misuse detection.

Keywords: Association Rules, Data Mining, Fuzzy logic, Genetic Algorithm, Intrusion Detection System.

I. INTRODUCTION

A. Data mining

Data mining [3] is also known 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.

B. 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:

1. 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.

Table 1: Various types of attacks described in four major categorizes

Denial of Service Attacks	Back, land, neptune, pod, smurf, teardrop, ping, apache, of death, back, mail bomb, UDP storm
Probes	Satan, ipsweep, nmap, portsweep , saint, mscan
Remote to Local Attacks	Ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster,xlock, guest, sendmail dictionary, xnsnoop
User to Root Attacks	Buffer_overflow, load module, Perl, root kit, xterm

C. Intrusion Detection System(IDS)

Intrusive activities to computer systems and networks are increasing due to the commercialization of the internet and local networks. The security of our computer systems and data is at frequent risks due to the widespread growth of the internet and increasing availability of tricks for intruding and attacking networks have made intrusion detection to become a critical component of network administration. An intrusion detection system watches networked devices and searches for malicious or doubtful behaviours in kinds of pattern in the audit data stream [8]. Basically it is the combination of software and hardware that attempts to perform intrusion detection and raise the alarm when possible intrusion happens

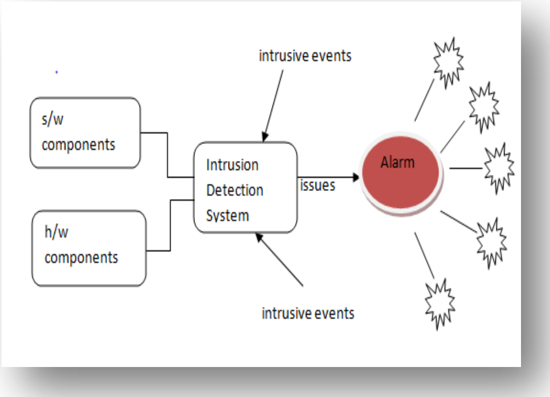

Figure 1: Block diagram of an IDS

D. Types of IDS

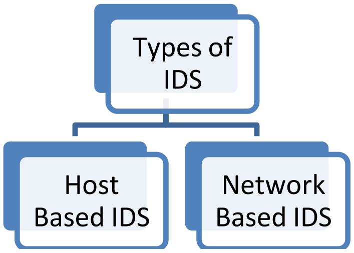

Figure 2: Types of IDS

a.*Host-based IDS*

HIDSs i.e Host based Intrusion Detection Systems evaluate information found on a single or multiple host systems, including contents of operating systems, system and application files. Basically Host based IDSs examine data held on individual computer that serve as hosts . In HIDS, software resides on each of the hosts that will be governing by the system hence it means it is an agent based system.

b. *Network-based IDS*

NIDSs i.e Network based Intrusion Detection Systems, evaluate information obtained from network communications, analyzing the stream of packets which travel across the network. As it not an agent-based system as compared to HIDSs , so low expense is brilliant advantage for NIDS because it is not necessary to install many monitoring systems

E. Detection Approaches

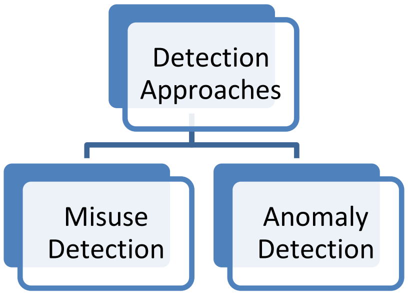

Figure 3: Different Detection approaches of IDS

a. *Misuse Detection*

It is also called as Signature based detection. It looks for patterns or user behaviour that matches known intrusions, which are stored as patterns or signatures. These hand coded signatures are laboriously provided by human experts based on their knowledge. If a pattern match is found, it signals an event then an alarm is raised. But the main drawback of this system is that it is unable to detect new or previously unknown intrusion. So only known intrusions will be caught up.

b. *Anomaly Detection*

It is also called as Profile based detection. Anomaly detection includes profiles (normal network behavior) which can be used to detect new patterns that deviate from the normal profiles. The main advantage of anomaly detection is that it may detect new intrusion that have not yet observed. A limiting factor of anomaly detection is the high rate of false positives.

II 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→Y) = P (X∪Y)
Confidence (X→Y) = P (Y|X)

Association rules were first developed to find correlations in transactions using retail data [8]. 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 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.

III. FUZZY LOGIC

Fuzzy Logic is derived from fuzzy set theory dealing with reasoning that is approximate rather than precise. Florez G. et al. [6] 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. [10]
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 [2]. 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 [3]. 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 [6]. 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.

IV.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. Any fuzzy association rule is in the following form:

If X is A then Y is B. (1)

In the above rule, X = {x1, x2, . . ., xp} and Y = {y1, y2, . . ., yq} are attribute sets. X and Y are disjoint subsets of I. A ={f *x*1 ,f *x*2 ,...,f *xp* } B={f_{y1},f_{y2}......f_{yp}) fuzzy sets associated with the corresponding attributes in X and Y. For example $f_{xk\,\epsilon}F_{xk}$ is a fuzzy set, defined on x_k domain. Each pair of (x_k , f_{xk}) is called an item, and each pair of (X, A) or (Y, B) is called an itemset. The first part of the rule ‘X is A’ is called the antecedent and ‘Y is B’ is called the consequent of the rule. The semantics of the rule is when ‘X is A’ is satisfied, we can imply that ‘Y is B’ is also satisfied. Here the word “satisfied” means there are sufficient amount of records which contribute their votes to the attribute/fuzzy set pairs and the sum of these votes is greater than a user specified threshold. An appropriate rule should have enough significance and a high certainty factor. Significance and certainty factor are two concepts, equivalent to sup-port and confidence .

V. 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 [19]. GA uses an evolution and natural selection that uses a chromosome-like data structure and evolve the chromosomes using selection, recombination and mutation operators [19]. 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. These positions could be referred to as genes. An evaluation function is used to calculate the goodness of each chromosome according to the desired solution; this function is known as “Fitness Function”. During the process of evaluation “Crossover” is used to simulate natural reproduction and “Mutation” is used to mutation of species [19]. For survival and combination the selection of chromosomes is biased towards the fittest chromosomes.

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 [19]. They are: i) the fitness function; ii) the representation of individuals; and iii) the GA parameters. The determination of these factors often depends on applications and/or implementation.

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 [9].

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 operator, new strings are created by exchanging information among strings of the mating pool. Types of crossover are explained in [7].

c) [MUTATION]: Mutation adds new information in a random way to the genetic search process [7]. 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 returns the best solution in current population.
6. [LOOP] Go to step 2.

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 [13].

VI. 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.

Table 2: Reviews of the techniques for Intrusion Detection

Technique	Review
Association rules	Association rule algorithm finds correlations between features or attributes used to describe a data set. But the existing Association rules cause the sharp boundary problem.
Fuzzy Association Rules	The basic concept to introduce the fuzzy set with association rules is to handle the sharp boundary problem in an appropriate way.
Genetic Algorithms	These algorithms are often used for optimization problems. When using fuzzy logic, it is often difficult for an expert to provide "good" definitions for the membership functions for the fuzzy variables. The genetic algorithms can be successfully used to tune the membership functions of fuzzy sets used by the intrusion detection system .

ACKNOWLEDGEMENT

I would like to thanks CSE department of RIMT-IET , Mandi Gobindgarh, Punjab.

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.

[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.

[11] Macros .M. Campos, Boriana L. Milenora, Creation & Deployment of Data Mining based Intrusion Detection Systems in Oracle Db 10g, in the *proceedings of 4th International Conference on Machine Learning & Applications*, 2005.

[12] Madjid Khalilian , Norwati Mustapha , Md Nasir Sulaim-an, Ali Mamat, Intrusion Detection System with Data Mining Approach: A Review, *Global Journal of Computer Science & Technology*, Volume 11 Issue 5 Version 1.0 April 2011

[13] Sathya s., Ramani R., Sivaselvi K., Discriminant Analysis based Feature Selection in KDD Intrusion Dataset, *International Journal of Computer Applications (0975 – 8887)*, Volume 31– No.11, October 2011.

[14] Naidu N. and Dharaskar R., An Effective Approach to Network Intrusion Detection System using Genetic Algorithm, *International Journal of Computer Applications (0975 - 8887)* ,volume 1 No.2, 2010.

BIOGRAPHIES

Harshna, pursuing M.Tech from RIMT-IET college, Mandi gobindgarh ,Punjab

Navneet Kaur , Assistant Profeesor in CSE department of RIMT-IET college , Mandi Gobindgarh , Punjab