MJRET
Open Access
Multidisciplinary Journal of Research in Engineering and Technology
Journal homepage: www.mjret.in
ISSN:2348 - 6953

INTRUSION DETECTION SYSTEM BASED ON FCA USING GA

Ms.Pallavi M. Shimpi[1], Prof. Vijay B. Patil[2]
Computer Science & Engineering Department[12]
MIT College
Aurangabad, India
pallavi_shimpi27@rediffmail.com[1],vijay.bpatil2006@gmail.com[2]

Abstract: *Intrusion detection is the process of supervising the events taking place in a computer system or network and analyzing them for signs of possible activities, which are assaulted or close to threats of violation for computer security.*

Incidents have many causes, such as malware (e.g., worms, spyware), attackers gaining unauthorized access to systems from the internet, and authorized users of systems who misuse their privileges or attempt to acquire extra privileges for which they are not empowered. Proposed method includes GA-based fuzzy Class Association Rule Mining with sub-attribute utilization and its application to classification, which can deal with discrete and continuous attributes at the same time. In addition, the proposed method is applied to both misuse detection and anomaly detection. Since the association rules used in the traditional information detection cannot effectively deal with alterations in network behaviour, it will better satisfy the real needs of abnormal detection to introduce the concept of fuzzy association rules to strengthen the adaptability. Experimental results with dataset KDD99Cup from MIT Lincoln Laboratory shows that the proposed method provides competitively high detection rate, i.e. 97% and PFR 3.45% compared with other machine-learning techniques.

Keywords: *Association Rule, Fuzzy Logic, Genetic Algorithm, IDS.*

1. INTRODUCTION

With the rapid development of network technology, the network computer system has become the intrusion target of hackers, network system security faces a huge threat, and intrusion detection technology becomes the hot topic in the field of network security [1]. IPS is a control tool, while an IDS is a visibility tool. IDS,monitors traffic at many different points, and provide visibility in security posture of the network. A good way is to compare an IDS with a protocol analyzer. A protocol analyzer is a tool that a network engineer uses to look detail into the network and observe what is going on. So An IDS can be a "protocol analyzer" for the security admin. The main objevtive of IDS is to reduce False Positive rate and to increase Detection rate. There are several ways to categorize an IDS: misuse detection vs. anomaly detection: in misuse detection, the IDS analyzes the information it gathers and compares it to large databases of attack signatures. Essentially, the IDS looks for a specific attack that has already been documented. Like a virus detection system, misuse detection software is only as good as the database of attack signatures that it uses to compare packets against. In anomaly detection, the system administrator defines the baseline, or normal, state of the networks traffic load, breakdown, protocol, and typical packet size.

The anomaly detector monitors network segments to compare their state to the normal baseline and look for anomalies. Network-based vs. host-based systems: in a network-based system, or NIDS, the individual packets flowing through a network are analyzed. The NIDS can detect malicious packets that are designed to be overlooked by a firewalls simplistic filtering rules. In a host-based system, the IDS examines at the activity on each individual computer or host. Passive system vs. reactive system: in a passive system, the IDS detects a potential security breach, logs the information and signals an alert. In a reactive system, the IDS responds to the suspicious activity by logging off a user or by reprogramming the firewall to block network traffic from the suspected malicious source

As an important supplement of the traditional prevention intrusion technology, intrusion detection is another fence to protect network computer systems. So to establish an effective and real-time intrusion detection system is a huge engineering task. After the emergence of a new class of invasion, intrusion detection system needs to real-time update the invasion match model, because in today world with increasingly high degree of information technology, even in a very short time delay, the new invasion would result in very large hazards. However, if only depends on the knowledge and experience of the system builders to analyze, classify the attack scene and system weak points, extract the characteristics of invasive means to store in the feature database, and then manually write the rules and models matched with the new invasion, if until when the feature the invasion monitoring

packets extracted is same to the characteristic of the database it will be judged as invasion, it is very likely that during this period of manual analysis and the preparation of the rules, the new invasion way has resulted in a significantly enough network disaster [2].

In such a system design process, human factors play a decisive role, because the system cannot adapt to the complex network environment, has not enough prevention for the endless new attack means of hackers, and the system self-adaptability and effectiveness of detection is extremely limited. Several Machine Learning algorithms, for e.g. Neural Network, Support Vector Machine, Genetic Algorithm, Fuzzy Logic, and Data Mining and more developed to detect intrusions for mix database. Various researches with data mining has been carried to find out new types of intrusions. There are some challenges for IDS: It looks like a defense tool which every organization needs but it also face some challenges like IDS technology itself is undergoing a lot of enhancements. It is therefore very important for organizations to clearly define their expectations from the IDS implementation.

IDS technology has not reached a level where it does not require human intervention, Of course today's IDS technology offers some automation like notifying the administrator in case of detection of a malicious activity, shunning the malicious connection for a configurable period of time, dynamically modifying a router's access control list in order to stop a malicious connection etc. But it is still very important to monitor the IDS logs regularly to stay on top of the occurrence of events. Monitoring the logs on a daily basis is required to analyze the kind of malicious activities detected by the IDS over a period of time. Today's IDS has not yet reached the level where it can give historical analysis of the intrusions detected over a period of time. This is still a manual activity. The IDS technology is still reactive rather than proactive. The IDS technology works on attack signatures. Attack signatures are attack patterns of previous attacks. The signature database needs to be updated whenever a different kind of attack is detected and the fix for the same is available. The frequency of signature update varies from vendor to vendor.

2. LITERATURE SURVEY

The survey, conducted explores the history of research in intrusion detection is performed in software in the context of operating systems in a single computer, a distributed system, or a network of computers. Various international conference papers, textbooks and Internet are the major source information considered under literature survey.

Since the first model for intrusion detection was developed by Dorothy Denning at SRI International, many intrusion detection systems, these systems are extremely diverse in the techniques they employ to gather and analyze data, most of them rely on a relatively general architectural framework, which consists of the following components:

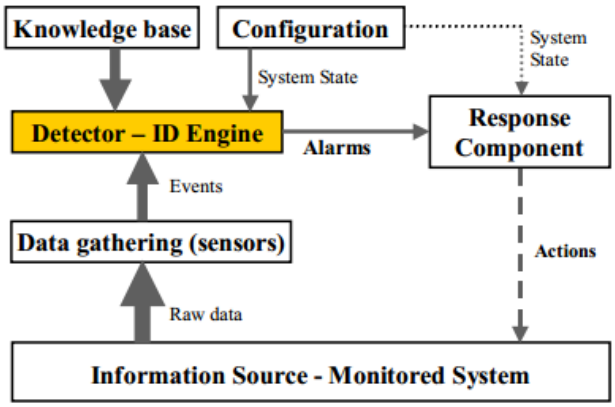

Figure 2.1: Basic architecture of intrusion detection system

2.1 Introduction

Information Systems and Networks are subject to electronic attacks. Attempts to breach information security are rising every day, along with the availability of the Vulnerability Assessment tools that are widely available on the Internet, for free, as well as for a commercial use. Tools such as SubSeven, BackOrifce, Nmap, L0ftCrack, can all be used to scan, identify, probe, and penetrate your systems. Firewalls are put in place to prevent unauthorized access to the Enterprise Networks. Let's, however, ask ourselves: Are the firewalls enough? Intrusion Detection System is one of the most important aspects of Computer/Network Security nowadays, though it has been introduced much earlier. Intrusion is not the only factor associated with breaching the network or computer security, but plays a major role to show its existence and dominance in rupturing the overall system security.

2.2 Computer Security

It is a technique for ensuring that data stored in a computer cannot be read or compromised by any individuals without authorization. Most computer security measures involve data encryption and passwords. Data encryption is the translation of data into a form that is unintelligible without a deciphering mechanism. A password is a secret word or phrase that gives a user access to a particular program or system. All Intrusion Detection Systems use one of two detection techniques:

- ***Statistical anomaly-based IDS***

An IDS which is anomaly based will monitor network traffic and compare it against an established baseline. The baseline will identify what is "normal" for that network- what sort of bandwidth is generally used, what protocols are used, what ports and devices generally connect to each other- and alert the administrator or user when traffic is detected which is anomalous, or significantly different, than the baseline. The issue is that it may raise a False

Positive alarm for a legitimate use of bandwidth if the baselines are not intelligently configured.

- ***Signature-based IDS***

A signature based IDS will monitor packets on the network and compare them against a database of signatures or attributes from known malicious threats. This is similar to the way most antivirus software detects malware. The issue is that there will be a lag between a new threat being discovered in the wild and the signature for detecting that threat being applied to your IDS. During that lag time your IDS would be unable to detect the new threat.

Proposed method uses fuzzy class association rule with genetic algorithm to detect intrusion from KDDcup Dataset,which gives 97% of Detection rate & 1.03% of False Positive rate.

3. METHODOLOGY

3.1 FCA with GA Approach

This approach uses genetic algorithm, fuzzy logic and class-association rule mining algorithm. Due to which, this proposed system works for both misuse and anomaly intrusion detection system.

3.2 Association rules:

As one of the most popular data mining methods for a wide range of applications, association-rule mining is used to discover association rules or correlations among a set of attributes in a dataset. The relationship between datasets can be represented as association rules. An association rule is expressed by $X \Rightarrow Y$, where X and Y contain a set of attributes. This means that if a tuple satisfies X, it is also likely to satisfy Y. The most popular model for mining association rules from databases is the a priori algorithm [8]. This algorithm measures the importance of association rules with two factors: support and confidence. However, this algorithm may suffer from large computational complexity for rule extraction from a dense database.

An association rules were originally developed as a tool for analysis of retail sales. A piece of sales data usually includes information about a transaction, such as transaction date and items purchased. Association rules can be used to find the correlation among different items in a transaction. For example, when a customer buys item A, item B will also be purchased by the customer with the probability of 90%. Agrawral and Srikant have presented some fast algorithms to mine association rules, including algorithm Apriori. Using the notation of Agrawal and Srikant let D = {T1, T2… Tn} be the transaction database with n transactions in total and I = {i1, i2… im} be the set of all the items where each if ($1 \leq j \leq m$) represents one kind of item. Then each transaction Tl ($1 \leq l \leq n$) in D records the items purchased, i.e., Tl I. Define an itemset as a nonempty subset of I.

An association rule will have the form: X ⇒ Y, c, s, where X I, Y I, X and Y are disjoint itemsets. Here is represents the support of this association rule and c represents the confidence of this association rule. Assume the number of transactions that contains both the itemset X and the itemset Y is n‘; support (X∪Y)=n‘/n) and c=support(X∪Y)/support(X).Intuitively, support(X) can be viewed as the occurrence frequency of the itemset X in the whole transaction database D, while c indicates that when X is satisfied, there will be the certainty of c that Y is also true .Two thresholds, minconfidence (representing minimum confidence) and minsupport (representing minimum support), are used by the mining algorithm to find all association rules X ⇒ Y, c, s, such that c ≥ minconfidence and s ≥ minsupport.

3.3 Integration of Fuzzy Logic with data mining

Although association rules and frequent episodes can be mined from audit data for intrusion detection, the mined rules or episodes are at the data level. Integrating fuzzy logic with association rules allows one to extract more abstract patterns at a higher level.

3.4 Mining Fuzzy Association Rules

Srikant and Agrawal have described a very popular algorithm for mining quantitative association rules that partitions quantitative attributes into different intervals.Unfortunately, a sharp boundary problem‖ results from using interval partitions. For example, suppose [1, 5] and [6, 10] are two intervals created on a quantitative attribute as shown in Figure 2.5. If the minimum support threshold is set at 30%, the interval [6, 10] will not gain enough support regardless of the large support near its left boundary, as shown in Figure 2.5. That is to say, although the value 5 has a large support and lies near the interval [6, 10], it will not make any contribution when counting the support of [6, 10].

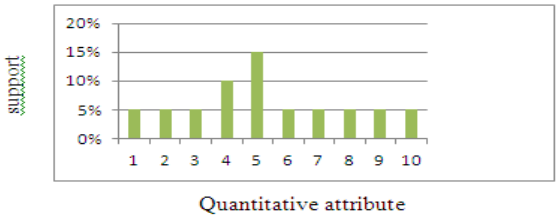

Fig 3.1: Example of Sharp Boundary Problem

In intrusion detection, the sharp separation of intervals may raise additional problems. For example, suppose the interval [1, 5] is mined as a normal pattern for the quantitative attribute. The values 6 and 10 will both be considered abnormal regardless of the difference in their deviations from the normal pattern. Likewise, a normal behavior that varies slightly from normal may fall outside the interval representing a normal pattern and be considered an anomaly. Similarly, an intrusion with a small variance may fall inside the interval and be undetected.

To address the sharp boundary problem, Kuok, Fu, and Wong have proposed to mine fuzzy association rules by using fuzzy sets to categorize a quantitative attribute. In the above example, the two intervals will be replaced by two fuzzy sets. Suppose the value 5 has a membership degree of 0.9 in the first set and 0.3 in the second set. Then it will contribute 0.9 to the support of the first fuzzy set and 0.3 to the second one. However, this means that the value 5 will be more important than other values since the sum of its contributions to different fuzzy sets has become greater than 1.

3.5 Overview of the rule mining based on GNP

A class-association-rule mining algorithm based on GNP has been proposed. In this section, the outline of GNP and its class association- rule mining is briefly reviewed [12].

GNP is one of the evolutionary optimization techniques, which uses directed graph structures instead of strings and trees. GNP is composed of three types of nodes: start node, judgment node, and processing node. Judgment nodes, J1, J2 , . . . , Jm (m is the total number of judgment functions), serve as decision functions that return judgment results so as to determine the next node. Processing nodes, P1, P2…. Pn (n is the total number of processing functions), serve as action/processing functions. The practical roles of these nodes are predefined and stored in the function library by supervisors. Once GNP is booted up, the execution starts from the start node, then the next node to be executed is determined according to the connection between nodes and a judgment result of the current activated node. Fig. 2.6 also describes the gene of a node in a GNP individual. NTi represents the node type such as 0 for start node, 1 for judgment node and 2 for processing node. IDi serves as an identification number of a judgment or processing node, for example, NTi = 1 and IDi = 2 represents node function J2. Ci1, Ci2…. denote the node numbers connected from node i. The total number of nodes in an individual remains the same during every generation. Three kinds of genetic operators, i.e., selection, mutation, and crossover, are implemented in GNP.

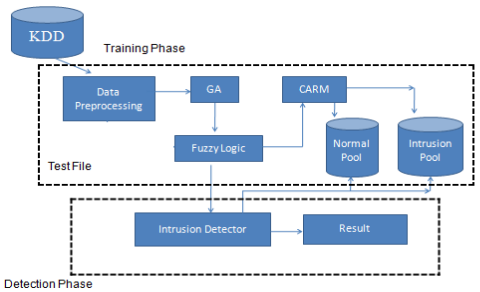

Figure 3.2: System Architecture

- ***Selection****:* Individuals are selected according to their fitness.

- ***Crossover:*** Two new offspring are generated from two parents by exchanging the genetic information.The selected nodes and their connections are swapped each other by crossover rate Pc.
- ***Mutation****:* One new individual is generated from one original individual by the following operators. Each node branch is selected with the probability Pm1 and reconnected to another node. Each node function is selected with the probability Pm2 and changed to another one.

3.6 Class Association Rules Mining

Association rules were originally developed as a tool for analysis of retail sales. A piece of sales data usually includes information about a transaction, such as transaction date and items purchased. Association rules can be used to find the correlation among different items in a transaction. For example, when a customer buys item A, item B will also be Agrawral and Srikant have presented some fast algorithms to mine association rules, including algorithm Apriori.[10] Using the notation of Agrawal and Srikant.[10] Let *I = {A1, A…, Al}* be a set of literals, called items or attributes. Let *G* be a set of tuples, where each tuple *T* is a set of attributes such that $T \subseteq I$. Let TID be an ID number associated with each tuple. A tuple *T* contains *X*, a set of some attributes in *I*, if $X \subseteq T$. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$. *X* is called antecedent and *Y* is called consequent of the rule. If the fraction of tuples containing *X* in *G* equals *x*, then we say that support(*X*) = *x*. The rule $X \Rightarrow Y$ has a measure of its strength called confidence defined by support $(X \cup Y)$ /support(*X*).

Calculation of $\chi 2$ value of rule $X \Rightarrow Y$ is shown as follows. Assume support(X) = x, support(Y) = y, support($X \cup Y$) = z, and the total number of tuples is N. We can calculate $\chi 2$ as

$$X^2 = \frac{N(z - xy)}{xy(1 - x)(1 - y)}$$

If $\chi 2$ is higher than a cutoff value, we should reject the assumption that X and Y are independent (3.84 at the 95% significance level or 6.64 at the 99% significance level).Let Ai be an attribute in a database with value 1 or 0, and k be class labels. Then, a class-association rule can be represented by $(Ap = 1) \wedge \cdot \cdot \cdot \wedge (Aq = 1) \Rightarrow (C = k)$ $k \in \{0, 1\}$ as a special case of the association rule $X \Rightarrow Y$ with fixed consequent C.Class-association rules satisfying the following are defined as important rules:

$$X^2 > \min X^2$$

$$\text{Support} \geq \text{minsupport}$$

Confidence ≥ minconfidence

Where χ2 min, supmin, and confmin are the minimum χ2, minimum support, and minimum confidence, respectively given in advance.

4.RESULTS

Computer systems and networks are often evaluated for the performance analysis through measurements, simulations, and emulations. Performance is one of the most important non-functional aspects of any (hardware, software) system. Performance evaluation applies certain techniques (measurements, analytical/simulation modeling) to existing or envisioned systems (in our case: IDS system, communication networks etc.) to assess performance measures of interest (delay, response times, throughput, jitter, processing times, etc.). The performance of developed system is directly measured in terms of whether intrusion is occurred or not. The testing is performed on the captured traffic from the network at the network interface, which was generated be various network elements, normally or intentionally. The effectiveness and efficiency of the proposed method are studied using DARPA98/99 database.

- ***Misuse Detection***

The proposed method for misuse detection is carried out with DARPA98 database in order to compare with other machine-learning methods. The training dataset contains 3342 connections randomly selected from DARPA98/99 database, among which 1705 connections are normal and the other 1637 connections are intrusion, where three types of attacks (nepture, smurf, and portsweep) are included. A total of 41 attributes are included in each connection; however, after the attribute division, 113 subattributes are assigned to the judgment functions in GNP. After 1000 generations, 3353 rules are extracted.

	Normal(T)	Intrusion(T)	Total
Normal(T)	183	4	187
Intrusion(T)	10	565	575
Total	193	569	762

Table 5.1: Testing result of the Probability Distribution with k = 0.5 in the misuse detection. DR, PFR, and NFR in the case of k = 0.5 are DR = (183+ 565)/762 = 98.16% PFR = 4/187 = 2.13% NFR =10/575= 1.73%

- ***Anomaly Detection***

The proposed method for anomaly detection is evaluated by the simulations with DARPA98 and DARPA99 database. The training database is intrusion free for the purpose of the anomaly detection. It contains 9137 normal connection records. After preprocessing, 30 attributes are included in every connection record. However, after the attribute division, 82

subattributes are assigned to the judgment functions in GNP. After 1000 generations, 5589 rules related to the normal connections are extracted.

	Normal(T)	Intrusion(T)	Total
Normal(T)	179	15	194
Intrusion(T)	8	571	579
Total	187	586	773

Table 5.1: Testing result of the Probability Distribution with k = 0.5 for anomaly detection. DR, PFR, and NFR in the case of k = 0.5 are,DR = (179 + 571)/773 = 97.02% PFR = 15/194 = 7.73%,NFR = 8/579= 1.38%

Finally, we can point out that the proposed method satisfies the desired features of network intrusion detection system which is applied on DARPA 98 and DARPA99 Dataset. The proposed method, that is, hybrid rule mining algorithm based on GNP, combines Fuzzy-based class association rule mining and probabilistic classification in order to extract more important rules from the database. In addition, the classification method is based on the probability distribution of the average matching degree between data and different class rules. As a result, simulations show higher DR and lower PFR, NFR, which means that Fuzzy based GNP class association rule mining has better performance than the conventional class association rule mining. The reason why the fuzzy data mining outperforms the crisp data mining is its characteristic of overcoming a sharp boundary problem. Fuzzy sets can help to overcome this problem by allowing a continuous attribute value to be a partial membership of more than one set.

5.CONCLUSION

Data mining methods are capable of extracting patterns automatically and adaptively from a large amount of data. Various methods related to intrusion detection system are studied and compared. Crisp data mining methods such as ADAM method, Random Forest algorithm are used for intrusion detection but suffer from sharp boundary problem which gives less accurate results. In proposed method, use of fuzzy logic overcomes the sharp boundary problem. Class-Association rules have been used to mine training data to established normal patterns for anomaly detection. An actual intrusion with a small deviation may match the normal patterns and thus not be detected. Therefore, integration of fuzzy logic with class-association rules and GA generates more abstract and flexible patterns for anomaly detection.

In this paper, we have proposed a GA-based fuzzy Class Association Rule Mining with Sub-Attribute Utilization and its application to classification, which can deal with discrete and

continuous attributes at the same time. In addition, this method was applied to both misuse detection and anomaly detection.

ACKNOWLEDGMENT

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them. I am highly indebted to Prof. Vijay B. Patil for his guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project I would like to express my gratitude towards my parents & member of MIT College for their kind co-operation and encouragement which help me in completion of this project.

REFERENCES

[1] Mabu S., Chen C., Shimada K., "An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming," IEEE Transactions Systems, Man, Cybernetics C, Application and Reviews, volume 41, number 1, pp. 130–139, January 2011.

[2] Hoque M., Mukit M. and Bikas M., "An Implementation of Intrusion Detection System using Genetic Algorithm," International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.2, March 2012.

[3] Lu W. and Traore I., "Detecting new forms of network intrusion using genetic programming," Computer Intelligence, volume 20, no. 3, pp. 474–494, 2004.

[4] Kaliyamurthie K., Parameswari D., Suresh R., "Intrusion Detection System using Memtic Algorithm Supporting with Genetic and Decision Tree Algorithms," IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 3, March 2012.

[5] Scarfone, Karen; Mell, Peter (February 2007). "Guide to Intrusion Detection and Prevention Systems (IDPS)". Computer Security Resource Center (National Institute of Standards and Technology)(800–94).http://csrc.ncsl.nist.gov/publications/nistpubs/800-94/SP800-94.pdf. Retrieved 1 January 2010.

[6] 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.

[7] Kddcup 1999data [Online]. Available: kdd.ics.uci.edu/ databases/kddcup99/kddcup99 .html.

[8] Han J., Kamber M., "Data Mining," Morgan Kaufmann Publishers, 2001.

[9] Shetty M. and Shekokar N., "Data Mining Techniques for Real Time Intrusion Detection Systems," International Journal of Scientific & Engineering Research Volume 3, Issue 4, April 2012.

[10] R. Agrawal and R. Srikant, —Fast algorithms for mining association rules,in Proc. 20th VLDB Conf., Santiago, Chile, 1994, pp. 487–499.

[11] Shingo Mabu, Member, IEEE, Ci Chen, Nannan Lu, Kaoru Shimada, and Kotaro Hirasawa, Member, IEEE, "An Intrusion-Detection Model Based On Fuzzy Class-Association-Rule Mining Using Genetic Network Programming," IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, vol. 41, no. 1, January 2011.

[12] D. E. Goldberg, Genetic Algorithm in Search, Optimization and Machine Learning. Reading, MA: Addison-Wesley, 1989.

[13] W. Li -"A Genetic Algorithm Approach to Network Intrusion Detection", SANS Institute, USA, 2004.

[14] Ajith A., Crina G., Yuehui C., —"Cyber Security and the Evolution of Intrusion Detection Systems", Information Management and Computer Security, 2005, Volume 9, No. 4, pp175-182.