IDS by Using Data Mining Based on Class-Association-Rule Mining and Genetic Network Programming

Mr. R. G. Raut. M. Tech (I.T.) Scholar
braut_rahul@yahoo.co.in

Abstract: Now a day's security is considered as major topics in networks, since the network has increasing widely day by day. Therefore, intrusion detection systems have paid more awareness, as it has an ability to identify intrusion accesses effectively. All these systems can spot the attacks and behave by trigger different errors .The proposed system includes a data mining method with fuzzy logic and class-association rule mining method which is based on genetic algorithm [1]. As the use of fuzzy logic, the recommend system can able to show the different type of features and also able to keep away from the different problems that are arising in to the suggested system approach. By using Genetic algorithm it is possible to find many rules and regulations and that are use to anomaly detection systems an association-rule-mining is very important technique that is used to find valuable rules and these rules are used by different users, instead of to find all the rules meeting the criteria that are useful for detection. Different results that are experimented with KDD99 [9] Cup database realise that the proposed approach gives more detection rates as compared to crisp data mining.

Keywords:- *Data Mining, Intrusion Detection System (IDS), Genetic Algorithm (GA), Network Security, Fuzzy Logic.*

*****

1. Introduction

Softness of any given network systems is playing a very key role in the security of the system as large amount of data and crucial information are going to be stored manipulated and used online.

Whatever be the IDS used today are signatures based that are developed by coding which one is manual and based on skillful knowledge. to known the different types of attacks these systems being connected to the different activity on the system that are going to be monitored.

The system using such approach facing one basic problem that is such system mainly fail to identify new attack rather than having the proper known signature.

Now days, to build detection models for IDSs more and more researcher increase their interest in data mining approaches. To detect unknown attacks this new system can use both attacks that are already known to the system and normal behavior of the system. The new system can produced in more quicker and automated way as compared to the conventional manually coding techniques.

Up tell now large number of data mining system for identifying intrusions had been developed, much of them could give better result as compared to the systems engineered by domain competence.

It is not just to design a well doing data mining system but it is required to generate the best and specific workable IDS and required to give full guarantee of the system that it can detect the intrusion best and system can be executed easily and it is very easy to use for the user and generalization.

For such system they have a large amount of difficulties specifically in the area of in the implementation and deployment of these systems. Generally there are three general categories of these difficulties: these are accuracy efficiency, and usability. As compared to the conventional signature based methods data mining-based IDSs having large amount of false positive rates which can make the system workable in the real environments

Now days, computer security become more and more primary issues. This is due to the emergence of electronic commerce, the tremendous use of computers and the rapid growth of computer networks.

When an attacker is trying to crack an information system or doing an action that is not allowed by legally this activity is called as an intrusion. Intrusion techniques may include systems password cracking, introducing software bugs in to system that is being configured, detecting untrusted traffic, or introducing the design flow of specific protocols. An IDS is a system for detecting intrusions and reporting to the proper authority.

Lastly to speak on these systems, to be able to deploy real time data mining-based IDSs, such system take large amounts of training data and as compared to the conventional approaches these are importantly more complex than traditional systems.

In order to design a beast data mining-based IDS it is required to refer these three group separately and also want to maintain a good balance between these group.

2.1 Problem Definition

Intrusion Detection System (IDS) is the most favorite technique used for finding the hackers attack in different way. At the time of solving the problem on the association rule create on the Genetic Network Programming (GNP).But in the research paper [4] to solve the problem. Other research problem is to combine the association rule and the fuzzy class.

2.2Research Problem Address

It has been proposed to develop the system such a

way that ,it can be the combition of both the GNP-based fuzzy Class Association Rule Mining and attribute like Utilization and its application to classification, that can deal with discrete and continuous attributes at the same time[1].At the same time such system can be used for misuses Detection and anomaly detection. For this system different Experiments were performed on practical data that are provided By KDD99 [9].

Whatever be the result produced by the system for misuse detection can provide more finding rate and low false positive rate, these are very important benchmark for security system that are being developed and For anomaly detection, the system provides very high detection rate and reasonable false positive rate even without advance knowledge of attack signatures, this is very important benefit of the system as compared to the conventional approach.

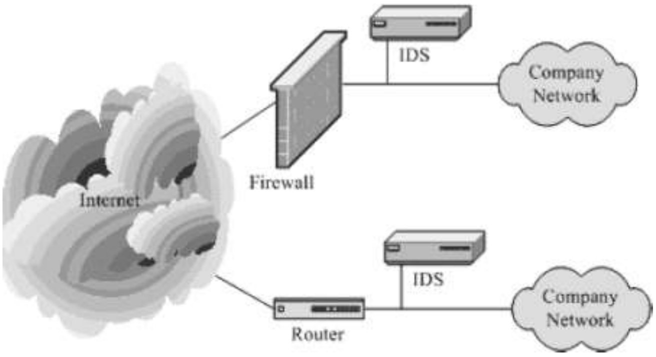

Figure 2.1: Intrusion Detection System

3. IDS

Intrusion Detection system now a days has been one of the most intrusting research area due to its capability of identifying various network attacks. Research manly point out the matter that concerned to develop the IDS based on data mining.

It found out the different approaches to mark three types of problems: precision, effectiveness and usability.

Data mining techniques are used to verify or analyze the audit data and to get features so that the it can differentiate the normal activities from intrusions so that it can acquire the better precision.

The calculation cost that are required for features generations are verified at high precision to acquire the better effectiveness

To achieve better usability, adaptive learning algorithms are used to facilitate model. To reduce the dependence on labeled data unsupervised anomaly detection algorithms are used.

3.1 The Project Idea

Once the attack or misuse shown the response can be automatic or manual and can be include ending of connection, or host of reaction directed to finding the attacker.

3.2 Purpose

Intrusion detection (ID) is a type of providing a security to the system for computers, networks and servers. An ID system collects the data and done the inspection of that

Data from different part of the computer system or a network to know the possible attacks that include both intrusions (attacks from outside the company) and misuse (attacks from within the company)

Different functions of IDS include:

- Track and getting the information of both user and system functions.
- To checking the system configurations and damages that are required to harm the computer system.
- Getting the access to the system and systems file
- Honesty
- Ability to identify attacks of typical patterns.
- Ability to identify the abnormal activity patterns of the computers systems.
- Ability to trace the user policy breaking methodology.

The purpose is to giving background information in the choosing and appraisal of Intrusion Detection Systems (IDS) and mainly to found out the areas for another research and to management of different technologies and Intrusion Detection and Response (IDR) systems. The new research has been added some important reports and monographs on intrusion detection system.

The main aim of research has to develop a good balance between the two sources, academic and the security and different hacking groups.

The main focus of the technology is to have the report that is on the Network IDS (NIDS) products and the known damages of the TCP/IP network protocols. The report gives details methods of detecting network intrusions and types in the development of Intrusion Detection Systems in the past years. It also trying to get the map the potential of secure networking against the availability of penetration tools and attacked source code from the Internet.

4.1Network based IDS (NIDS)

Intrusion detection system is a hardware or software that can be used to spot out the harmful likewise DOS, Port scanning trying to crack the system. All the incoming packets are scanned and harmful patterns are found out by the NIDS.

Different ports that are required by TCP connection need to be scanned, it may be consider by someone that it is required to conduct the port scan of all system in given network. Mostly all the system trying to scan the incoming data or information as in the same manner that conventional system can do.

In contrast to conventional system what exactly the NIDS can done? It can done the scanning of all the incoming network traffic more over it can scan the important information that are given out side to the system.

4.2 HIDS

Host-based IDS can scan all internal parts of the computer system and the position of those parts.

Another main function of HIDS are what resources are accessed by which program and discover these resources for example verification and validation process of the system can start to modifying the system password from data base at regular interval of time.

What is the current position of a system? To scan it this is again another important function of the HIDS, This research describe a data mining approach as well as signature based for adaptively building Intrusion Detection (ID) models. The central idea of the approach is to let signature based agent detect old attacks and the anomaly based new attacks by using audit programs to extract large scale set of features that describe each network connection or host session.

From past time Researchers have used many techniques for developing the latest (IDS) but still we do not have an effective IDS. [12] For this it is required to add different approaches of data mining and different approaches of expert systems. This type of integrated system can be covered more area and try to detect intrusion more effectively.

The audit data of the system can be found for consistent and useful patterns and after that take these data as normal behaviours in profiles of the system

The deviation in the system can be found out by the expert system and also trying to give the alarm for such a deviation in the system that are happened due to the intrusion in the system.

The work of the IDS can be judgement carrying out some time due to this importance of the work is justified. [13]

Research mainly point out some topics that are concerned to developing a data mining-based IDS using genetic programming. It gives details of the approaches to address three types of above discussed issues.

5.1Proposed Approach

The research has two parts where each of these parts works in a different position. The first stage consider the training stage GA and fuzzy-association rule mining algorithm and a set of classification rules are created from KDD dataset. The Second part consider intrusion detection stage, the generated rules are used to classify incoming data from a test file. Once the

Rules are created. Intrusion Detection is simple and efficient.

5.2 Data Pre-processing

For scanning the incoming and outgoing data traffic can be scanned by using the process called data pre-processing. This is not a single process but it is collection of different processes like creating a different data set of incoming data base, cleaning that data, integrating the data set, constructing the features, and

The data pre-processing is incorporation of steps like creation of dataset, data cleaning, integration, feature construction, organizing the different data tables and data base and selecting the different feature. It is very instructing to see that it can take overall process attempt near about 50.particular data set used in the data pre-processing is a is grouping of different element and that can describe the large amount of quality of the data The given elements may be in quality or quantity. In nature having different values in different rang. Data scanning process may have some effect on the character and the values of elements. Elements having the big values can govern the elements with small value. This process is used for removing such governance by all the data are scald within a particular range. The element that are in particular quantity can be easily normalized. but in case of large number of data element formal range are transformed in to numeric range. It is required to apply some condition for assigning the numerical values to the attributes. By applying this method when the large no of data attributes are transformed in to the particular data attributes, the data pre-processing can be applied to them.

Steps for pre-processing of attributes/features are shown in the following algorithm:

Algorithm: Grouping the
KDD dataset, Feature
[9] extraction. Input:
KDD dataset

Output: Dataset into two classes i.e. rule pool (Normal And attack)

1. Select KDD dataset [9]
2. Transform elements to numeric value
3. Find maximum value for each elements/feature
4. Select important elements/features
5. Store rules in rule pool

In above algorithm, classification method data mining is used for classifying the whole dataset into two classes i.e.”normal” and”attack”. Selection of feature is necessary because the use all available features are computationally not possible.

5.2.1 Fuzzy Logic

By using the advantage of fuzzy theory for keeping the every continuous attribute value in [0, 1] Complex system can be described in many linguistic approach by using fuzzy concept

Data base attribute can be converted in to five Linguistic terms (Vey low, Low, Middle, High, Very high).Using this five linguistic terms for single constant Attribute, it gives more perfect membership value that is related to the constant attribute.

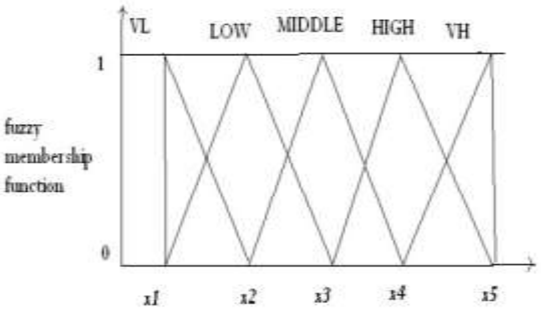

Figure 5.1: Fuzzy Membership functions

The similar structural terms are identified by the combination of both the membership function of trapezoidal and triangular. With the solution of GNP

The variables like x1, x2, x3, x4 and x5 are also solved. Constant attribute has its own membership worth. The variable value of fuzzy membership function should be adjusted during development process. In the association rules Membership values

of fuzzy are used to find the changes in given GNP. Following table shows example of small database with two constant attributes

Sr.no	A1	A2
1	100	10000
2	200	8000
3	300	6000
4	400	4000
5	500	2000

Table 5.1: small database example

5.3 Genetic Network Programming

Data pre-processing algorithm generates regulations that are going to fixed in the pool that is called as rule that has Normal rule pool contains normal records and attack rule pool contains records for intrusion. The following algorithm is common for both i.e. normal and attack rule pool and explains about the genetic algorithm and its operators.

Rule pool generate using genetic network programming concepts

Input: Pre-processed dataset generation no. (G), and size of given population.

Output: maximum number of rules in the given pool of rule

1. Initialize the population
2. N is population size, T (value of strength fn.= 0)
3. User input for number of generations (K)
4. Initialization of given independent (A) = 1
5. Initialize the strength counter (B) = 1
6. Select two chromosomes (or rules) from population
7. Increment A by 2, S by 1
8. Crossover operator are applying to the chromosome
9. Changing operator are applying to the chromosome
10. If rule is present in rule pool then go to step 4 for next rule
11. Else

Measures the connections Ntc find by r rule

Measures then no of relationship in the data set that are to be trained Nt

Measures e the no of usual relationship Nni wrongly find by r rule

Measures the no of normal relationship in the data set that are to be trained Nn

6.1 Class-Association-Rule mining (CARM) Algorithm

Statement of class association-rule mining is defined as give R = B1, B2, B3… B1is a set of error known as parts/elements

Give K be a set of data parts, where each data part T is defined as a element set in such way that $T \subseteq R$.

Suppose to Give a PID is an ID number and that are concerned with each and every data parts of the data set.

A particular data parts called as a T contains M, be the some element in the data base called as R, if it can be taken as a $M \subseteq T$.

An association rule can be defined as it is an indication of the form $M \Rightarrow N$, where $M \subset R$, $N \subset R$, and $M \; ^{n} \; N = \varphi$. Where,

M is known as predecessor and N is known as accordingly of that regulation are used in the process. If the piece of that data parts contains N within K equals to n, then it is says that concerned (M) = m.

M^2 can be calculated as the value of the given rule $M \Rightarrow N$ is given by. Here it required to assume some Assumption as the term (M) =m and the term (N) = n, then the result that are taken as the term consider here is (M U N) = O, then the total no of data parts are equal to S

6.2 GNP-Based Class-Association Rule

The basic function of scanning the element value in given data parts of the data set is performed by some judgment node and it can present in the GNP.

For the starting of the class association rule processing node P1 is used for this function, A particular element that can be taken here are B1,B2,B3 and that are consider as a judgment function and whose value can be taken as one.

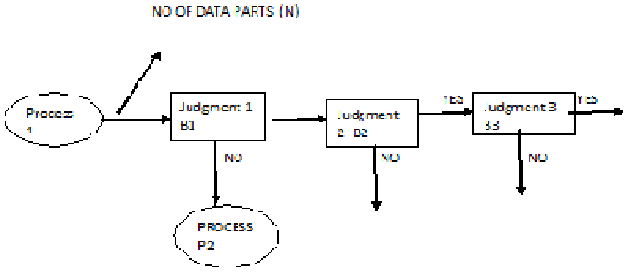

Figure 6.1 State Transmission

Suppose the given data parts can satisfy the given condition of the judgment function it required to take Yes-side curve as shown in fig.6.1 after that what is the condition of the next judgment function is inspected for finding the more rules and regulation.

One important point need to remember here is there is no any way that can attaches to the another node called as P2 for checking the different rules and regulation of another data set.

For example, the class-association rules such as

$(X = 1) \Rightarrow (K = 1)$

$(X = 1) \wedge (Y = 1) \Rightarrow (K = 1)$

$(X = 1) \wedge (Y = 1) \wedge (Z = 1) \Rightarrow (K = 1)$ $(X = 1) \Rightarrow (C = 0)$

$(X = 1) \wedge (Y = 1) \Rightarrow (K = 0)$

$(X = 1) \wedge (Y = 1) \wedge (Z = 1) \Rightarrow (K = 0)$

Data parts are inspected by using the following point. Whatever may be the first given data base is need to be read and then state transmission process start from node P1. Suppose in this process positive approach is taken then it can be marked as YES and ,at that time which is the current node is taken that can be transform in to the state that is called as judgment node. and suppose negative approach is taken then this can be marked as NO and then state can be transformed from current state to state P2 for discovering another regulation.

The above mentioned processes is going to be repeated until the process can be reaches to the node called as Pn.

Lastly, all the data parts are checked by repeating the above mentioned steps.

One rule is need to follow in this process that is no of judgment function must be equal to the no of element in the given data base.

6.3 Association Rules Measurement

For total no of data parts N is consider here. And again another consideration is x, y and z be the no of data parts that are transformed to YES side to the J node called as J1,J2 and J3 .

Mainly in this process the node called as processing from which this whole process can be started are responsible for counting the numbers and taking the measurement.

Taking an example, in the first rule case $(x = 1) \Rightarrow (k = 1)$, the concerned is x(1)/N and the credence is x(1)/x. In the second rule case $(X = 1) \wedge (Y = 1) \wedge (Z = 1) \Rightarrow (K = 1)$, the concerned is z(1)/N and the credence is z(1)/c.

7. Ranking

Intrusion are ranked mainly in two parts

7.1 Misuse

It is one concept to detecting a computer attacks or network attacks. in this concept at first abnormal system behavior is defined after that any system behavior is defined that is called normal behavior

This type of detection also been used to refer all kinds of computer misuse.

7.2 Anomaly

Anomaly based IDS is system for detecting the system misuse and intrusion by monitoring the system different activity and after that classify it in to normal or anomalous .this ranking is depend on some rules and trying to find all misuse that are consider as it is not a normal system operation.

8.1 System Architecture

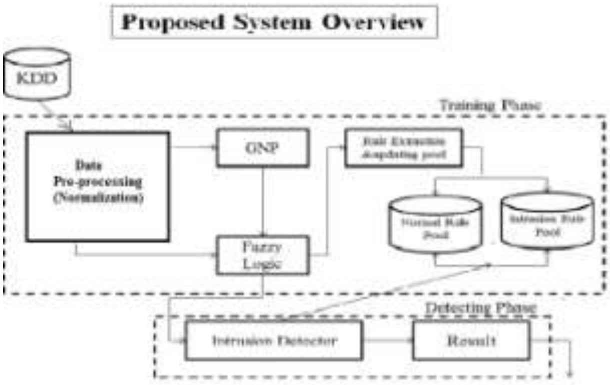

Figure 7.1: Proposed System

8.2 KDD

Knowledge Discovery in Databases" or KDD is an combination of more than one sub branch of computer science, is the computational action of finding design in large data base that can also involve some another methods like it is the addition of different machine learning ,data base and neural network and artificial intelligence. The primarily goal of the data mining concept is getting proper information from a large data base and convert it into an better and comprehensibly format for future use. Beside from the raw inspection methods, it include database and data management concept, data pre-processing, model and inference considerations,complesity,postprocessisng,metrics, visualization.

8.3 GNP

For scanning the given data base some attribute are required and that essential element must be present in GNP.It can be made up from different states mainly called as judgment or processing state and all the states in GNP are connected to each other by using directed graph. This type of graph structure enables to the GNP to reusing the different states in given data base.

8.4 Fuzzy Logic

In this part the advantages of fuzzy logic are taken to minting the range of attributes in between 0 and 1again it can have a fixed value in this range also.

8.5 Training Phase

8.5.1 Rule Pool

Data pre-processing algorithm generates rules which are stored in the pool that is called as rule and they are having Normal rule pool contains normal records and attack rule pool contains records for intrusion. One algorithm is common for both i.e. normal and attack rule pool and explain out the genetically algorithm and its operators.

8.6 Detecting Phase

8.6.1 Intrusion Detector

At the end of this algorithm large number of regulation will be available for further processing. For Anomaly detection, the quantity of regulation matters more than quality, whereas for misuse detection quality rules are required. So for both detection systems this algorithm is best suited.

8.6.2 Result

Finally the designed system detects the anomaly and misuse concepts with greater efficiency, accuracy and usability as compared to the previous system

9 References

[1] Shingo Mabu,Ci Chen, NannanLu, Kaoru Shimada ,and Kotaro Hira- sawa," 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

[2] Swati Dhopte, N. Z. Tarapore,"Design of Intrusion Detection System using Fuzzy Class-Association Rule Mining based on Genetic Algorithm" International Journal of Computer Applications (0975 -8887) Volume53-No.14,September2012.

[3] Jonatan Gomez and Dipankar Dasgupta," Evolving Fuzzy Classifiers for Intrusion Detection" Proceedings of the 2002 IEEE Workshop on Information Assurance United States Military Academy, West Point, NYJune2001.

[4] Zohair Ihsan, Mohd Yazid Idrisand Abdul Hanan Abdullaha," Attribute Normalization Techniques and Performance of Intrusion Classifiers: A Comparative Analysis". Life SciJ 2013;10(4): 2568-2576] (ISSN:1097-8135).

[5] Mohammad Sazzadul Hoque, Md. Abdul Mukit and Md. Abu NaserBikas," AN IMPLEMENTATION OF INTRUSION DETECTION SYS- TEM USING GENETICAL

[6] Nivedita P. Chaudhari and Dr.Leena Ragha,"SMART NETWORK INTRUSION DETECTION SYSTEM USING HYBRID APPROACH" Inter- national Journal of Research in Advent Technology(IJRAT)Vol.1,No. 2, August2013,

ISSN: 23219637.

[7] Processes,"presented at the IEEE Symp. Secur. Privacy,LosAlami- tos,CA, 1996.

[8] Kddcup 1999data[Online]. Available: kdd.ics.uci.edu/ databases/kd- dcup99/ kddcup99.html.

[9] S.Manganaris,M. Christensen, D. Serkle, andK. Hermix,"A data mining analysis of rtidalarms,"presented at the2ndInt. Workshop Re- centAdv.IntrusionDetect., West Lafayette, IN,1999.

[10] D. E. Denning, "An intrusion detection model," IEEE Trans. Softw. Eng.,vol. SE-13, no. 2,pp. 222-232, Feb. 1987.

[11] J. R. Koza, Genetic Programming, on the Programming of Comput- ers by Means of Natural Selection. Cambridge, MA:MITPress,1992.

[12] J.R. Koza, Genetic Programming II, Automatic Discovery of Reusable Programs.Cambridge, MA:MITPress,1994.

[13] Mabu,Chen,NannanLu,ShimadaandHirasawa,IEEEa nsactions On Systems, Man, And Cybernetics-PartC: Applications And Reviews, Vol.41, No. 1, January2011

Mr.R.G.Raut
B.E.(I.T.),M.Tech(I.T.) Scholar
Bharati Vidyapeeth Deemed University,
College of Engineering, Pune, MH, India.
braut_rahul@yahoo.co.in