Cluster Based detection of Attack IDS using Data Mining

[1]Manisha Kansra, [2]Pankaj Dev Chadha
[1]Research scholar, [2]Assistant Professor,
[1]Department of Computer Science Engineering
[1]Geeta Institute of Management and Technology , Kurukshetra, India

***Abstract* - One of the most important challenges to intrusion detection are the problem of misjudgment, misdetection and lack of real time reaction to the attack. In the recent years, as the second line of defense after firewall, the intrusion detection technique has got fast development. a mixture of data mining techniques such as clustering, categorization and association rule detection are being used for intrusion detection. This research proposed IDS using by integrated signature based (Snort) with abnormality based (Naive Bayes) to enhance system security to detect attacks. This research used Knowledge Discovery Data Mining (KDD) CUP 20 dataset and Waikato Environment for Knowledge Analysis (WEKA) program for testing the proposed hybrid IDS.**

***Index Terms* - IDS, Attack. Traditional IDS,J48**

I. INTRODUCTION

The intrusion detection systems (IDS)[1] using the traditional methods are limited in detecting unknown intrusion behaviour or updating the profile in real time and the maintenance will be slow and heavy, also the rate of false positives and false negatives is high compared to IDS using data mining approach, further limits of traditional IDS listed in. Thus several research focus to resolve these issues by using the data mining technology into the IDS, which makes them automatically produce accurate detection model from a large number of audit data.

Data mining is a set of techniques and methods in the field of statistics, mathematics and computer science to extract, identify valid, novel, potentially[1] useful, and ultimately understandable patterns in massive data. Various data mining's methods and algorithms have been used such as classification tree and support vector machines for intrusion detection,Genetic Algorithms, Neural Networks, and Clustering, all these methods helps to provide a good level of security to the systems from external and internal attacks, also from new attacks. Detecting attacks is an essential need in networks. The dataset used for network anomaly detection well-known as KDD Cup 1999.

1.2 Traditional IDS:-There are two types of traditional intrusion detection system:

- Anomaly Detection - It refers to detect abnormal behaviour of host or network. It actually refers to storing features of user's usual behaviours hooked on database, then its compare user's present behaviour with database. If there occurs a deviation huge enough, it is said that the data tested is abnormal. The patterns detected are called anomalies. Anomalies are also referred to as outliers.
- Misuse Detection - In misuse detection approach, it defines abnormal system behaviour at first, and then defines any other behaviour, as normal behaviour. It assumes that abnormal behaviour and activity has a simple to define model. It advances in the rapid of detection and low percentage of false alarm. However, it fails in discovering the non-pre-elected attacks in the feature library, so it cannot detect the abundant new attacks.

1.1.1 Types of IDS:-

- Host Based IDS - It refers to intrusion detection that takes place on a single host system. It gets audit data from host audit trails and monitors activities such as integrity of system, file changes, host based network traffics, and system logs. If there is any unlawful change or movement is detected, it alerts the user by a pop-up menu and informs to the central management server. Central management server blocks the movement, or a combination of the above two. The judgment should be based on the strategy that is installed on the local system.
- Network Based IDS - It is used to supervise and investigate network traffic to protect a system from network-based threats. It tries to detect malicious activities such as denial-of-service (Dos) attacks and network traffic attacks. Network based IDS includes a number of sensors to monitors packet traffic, one or more servers for network management functions, and one or more management relieves for the human interface.
- Hybrid Intrusion Detection - The recent development in intrusion detection is to combine both types host-based and network-based IDS to design hybrid systems. Hybrid intrusion detection system has flexibility and it increases the security level. It combines IDS sensor locations and reports attacks are aimed at particular segments or entire network.

1.1.2 Drawbacks of traditional IDS:-Intrusion Detection Systems (IDS) have many drawbacks due to the fast development of technology. The drawback of the current and traditional IDS are:

- Threshold detection - certain attributes of user and system behaviour are expressed in terms of counts, with some level established as permissible. Such behaviour attributes can include the number of files accessed by a user in a given period

of time, the number of failed attempts to login to the system, the amount of CPU utilized by a process. Use this technique in AnomalyBased Intrusion Detection System generate a high level of false positives alarms.

- False positives - a false positive occurs when normal attack is mistakenly classified as malicious and treated accordingly. The solution is to investigate and review the IDS configuration to prevent the false positive from occurring again.
- False negatives - A false negative occurs when an attack or an event is either not detected by the IDS or is considered benign by the analyst. Ordinarily the term false negative would only apply to the IDS not reporting an event.
- Updates lag - the main issue occurs to Signature-Based Intrusion Detection System is the update lag. In other words, will be always a lag between the appearance of new thread and the IDS's updates.
- Data size - the amount of data the analyst can efficiently analyze.

II,Data Mining Assists for intrusion detection

The central theme of intrusion detection using data mining approach is to detect the security violations in information system. Data mining can process large amount of data and it discovers hidden and ignored information. To detect the intrusion, data mining consist of following process like classification, clustering, association rule learning and regression. It monitors the information system and raises alarms when security violations are founded.

2.1 Neural Networks:- Neural Network was traditionally used to refer a network or biological neurons. In IDSs neural network has been used for both anomaly and misuse intrusion detection. In anomaly intrusion detection the neural networks were modeled to recognize statistically significant variations from the user's recognized behaviour also identify the typical characteristics of system users. In misuse intrusion detection the neural network would collect data from the network stream and analyze the data for instances of misuse. In neural network the misuse intrusion detection can be implemented in two ways. The first approach incorporates the neural network component into an existing system or customized expert system. This method uses the neural network to sort the incoming data for suspicious events and forward them to the existing and expert system. This improves the efficiency of the detection system. The second method uses the standalone misuse detection system. This system receives data from the network stream and analyzes it for misuse intrusion. It has the ability to learn the characteristics of misuse attacks and identify instances that are unlike any which have been observed before by the network. It has high degree of accuracy to recognize known suspicious events. Generally, it is used to learn complex non-linear input-output relationships.

2.2 Fuzzy Logic:- Fuzzy logic is derived from fuzzy set theory it uses the rule based systems for classification. Fuzzy can be thought of as the application side of fuzzy set theory dealing with sound thought out real world expert values for a complex problem. The fuzzy data mining techniques used to extract patterns that represent normal behaviour for intrusion detection. The sets of fuzzy association rules are used to mine the network audit data models and to detect the anomalous behaviour the set of fuzzy association rules are been compared to identify the similarity. If the similarity values are below a upper limit, an alarm raises.

2.3 Bayian Classifier:- A Bayesian Classifier provides high accuracy and speed for handling large database. In network model Bayesian classifier encodes the probabilistic relationship among the variable of interest. In intrusion detection this classifier is combined with statistical schemes to produce higher encoding interdependencies between the variables and predicting events. Bayesian belief networks based on the joint conditional probability distributions. The graphical model of casual relationships performs learning technique. This technique is defined by two components-a directed acyclic graph and a set of conditional probability tables.DAG represents a random variable these variables may be discrete or continuous. For each variable classifier maintains one conditional probability table (CPT).It require higher computational effort.

2.4 K-Nearest Neighbour:- K-Nearest Neighbour (k-NN) is a type of Lazy learning, it simply stores a given training tuple and waits until it is given a test tuple. It is an instance based learner that classifies the objects based on closet training examples in the feature space. For a given unknown tuple,a k-Nearest neighbour looks the pattern space for the k-training tuples that are closest to the unknown tuple.It is the simplest algorithm among all the machine learning algorithms. Here the object is classified by a majority vote of its neighbours. The object is simply assigned to the class of its neighbour only in the case of K=1.For a target function this algorithm uses all labeled training instances model. To obtain the optimal hypothesis function algorithm uses similarity based search. The intrusion is detected with the combination of statistical schemes. This technique is computationally expensive and requires efficient storage for implementation of parallel hardware.

2.5 Decision Tree:- Decision tree is a classification technique in data mining for predictive models. Decision tree is a flowchart like tree structure where internal node represents a test on attribute, branch represents an outcome of the test and leaf node represents a class label. From the pre classified data set it inductively learns to construct the models. Here each data item is defined by the attribute values. Initially decision tree is constructed by set of pre-classified data. The important approach is to select the attributes, which can best divide the data items into their respective classes based on these attributes the data item is partitioned. This process is iteratively applied to each partitioned subset of the data items. If all the data items in current subset belongs to the same class then the process get terminate. Each node contains the number of edges, which are labelled along with a possible value of attribute in the parent node. An edge connects either a node or two nodes. Leaves are always labeled with a decision value for classification of the data. To classify an unidentified object, the process is starts at the root of the decision tree and follows the branch. Decision trees can be used for misuse intrusion detection that can learn a model based on the training data

and predict the future data from the various types of attacks. It works well with large data sets. Decision tree model also be used in the rule-based techniques with minimum processing. It provides high generalization accuracy.

III.Purposed System

Due to the increase of internet technology in the past few years network traffic has also been increased to a great extent. Data travelling over the network has become a hot topic for the researchers because security is concerned for this data. Intrusion is the activity that violates the security policy of the system. Actually it is a deliberate unauthorized attempt to access and manipulate information.

Intrusion detection is a process which is used to identify the intrusion, and is based on the belief that the intruder behavior will be significantly different from the lawful user. Intrusion Detection System (IDS) are usually deployed along with other defensive security mechanisms, such as firewall and verification, as a succeeding line of defense that protects information systems.

3.1 .J48 decision tree classifier*:*

J48 is the decision tree based algorithm and it is the extension of C4.5. With this technique a tree is constructed to model the classification process in decision tree the internal nodes of the tree denotes a test on an attribute, branch represent the outcome of the test, leaf node holds a class label and the topmost node is the root node. Model generated by decision tree helps to predict new instances of data [3].

Algorithm [1] J48:

```
INPUT
D // Training data
OUTPUT
T // Decision tree
DTBUILD (*D)
{
T = Null;
T = Create root node and label with splitting attribute;
T = Add arc to root node for each split predicate and label;
For each arc do
D = Database created by applying splitting predicate to D;
If stopping point reached for this path, then
T'= Create leaf node and label with appropriate class;
Else
T' = DTBUILD (D);
T = Add T' to arc;
```

3.2 **Naïve Bayesian classifier:**

Bayesian classification represents a supervised learning method as well as statistical method for classification. It is simple probabilistic classifier based on Bayesian theorem with strong independence assumption. It is particularly suited when the dimensionality of input is high. They can predict the probability that a given tuple belongs to a particular class. This classification is named after Thomas Bayes (1702-1761) who proposed the bayes theorem. Bayesian formula can be written as [4]: $P(H \mid E) = [P(E \mid H) * P(H)] / P(E)$ The basic idea of Bayes's rule is that the outcome of a hypothesis or an event (H) can be predicted based on some evidences (E) that can be observed from the Bayes's rule.

IV.Simulation Results

Experiment are performed on Weka with 10 fold cross validation. Ten fold cross validation has been proved to be statistically good enough in evaluating the performance of the classifier The first step is to find the number of instances of diabetes dataset using both Naïve Bayes and j48 classification algorithm. In the next step experiment calculates the classification accuracy and cost analysis. Confusion Matrix**: -** Confusion matrix contain information about actual and predicted classification. Standard terms defined for this matrix .

True positive –if the outcome of prediction is p and the actual value is also p than it is called true positive(TP).

False positive-if actual value is n than it is false positive(FP)

Precision – precision is measure of exactness and quality

$$\text{Precision} = tp/(tp + fp)$$

Recall- measure of completeness and quantity

$$\text{Recall} = tp / (tp + fn)$$

Step1 Now we find out Centroid,We select weka 3.6.2 in cluster.

Fig4.1 : Select Cluster

RESULT FOR CLASSIFICATION USING J48 J48 is a module for generating. When appplying J48 on diabetes dataset result are as given below

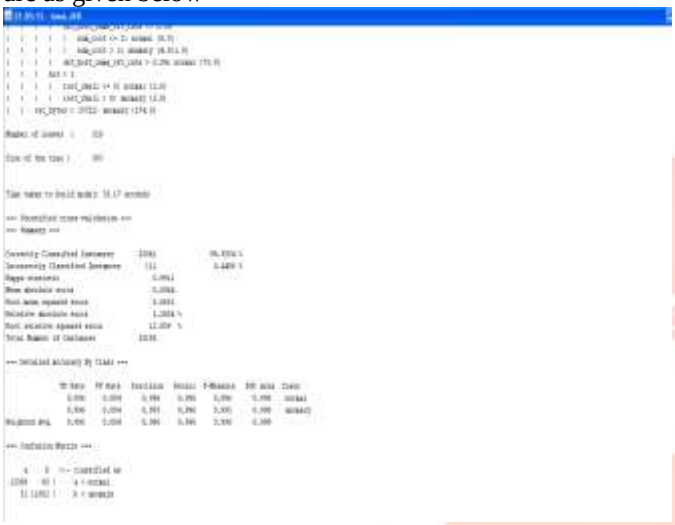

Figure 4.2: Confusion matrix for J48

Figure 4.3 : Cost analysis of J48 for class Normal

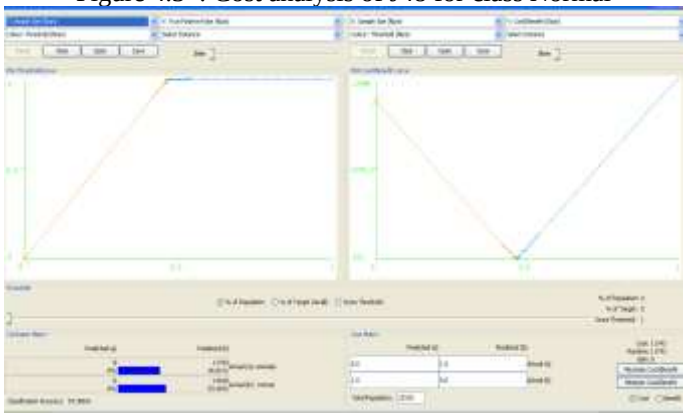

Figure4.4 : Cost analysis of J48 for class Anomaly

RESULT FOR CLASSIFICATION USING NAÏVE BAYES

When Naïve Bayes algorithm is applied on diabetes dataset, we got the result shown as below on figure .

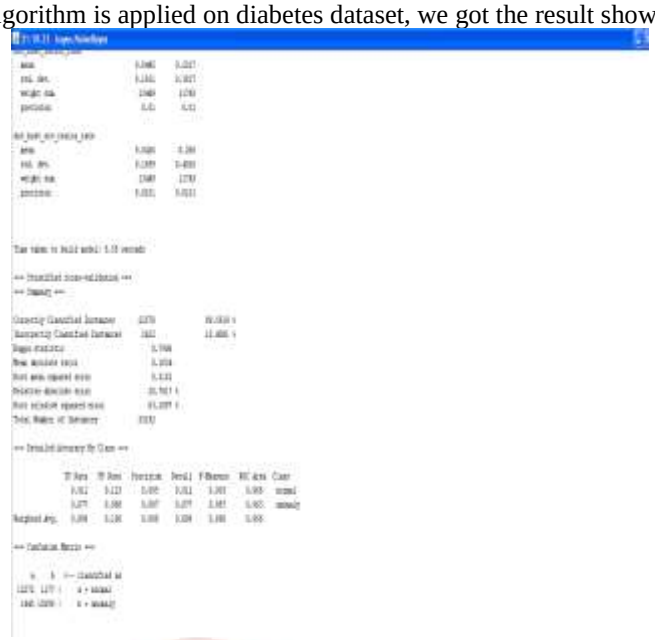

Figure 4.5: Confusion matrix for native bays

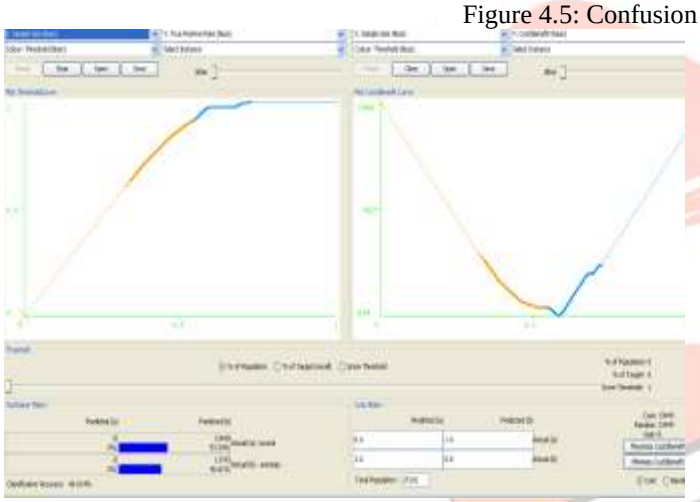

Figure4.6 : Cost analysis of Native Bays for class Normal

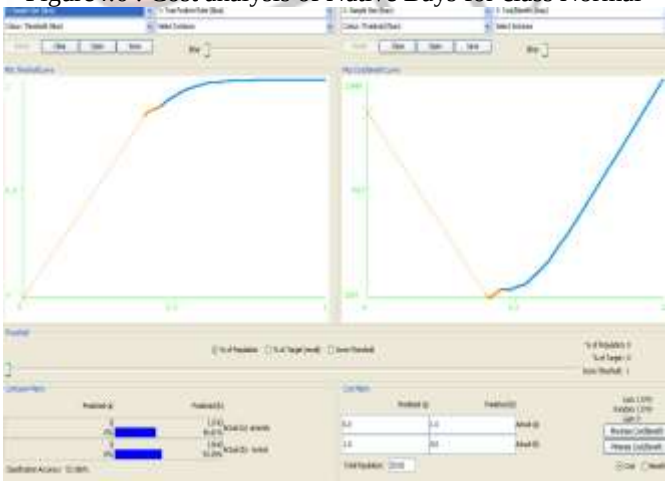

Figure4.7 : Cost analysis of Native Bays for class anomaly

V.Conclusion

The preliminary Experiments with the KDD Train 20 percent audit data have shown that this approach is able to successfully detect intrusive program behavior. According to the results and comparisons have been done in this research between all the even uses each of the algorithms Naïve Bayes, J48graft and Bayes Net, in term of accuracy, detection rate, false alarm rate and taken time to build model. In some strategies such as Percent Correct Classification (PCC) accuracy, the percentage of correctly

classified instances in J48graft was higher than percentage of correctly classified instances with uses bayes net and naïve bayes respectively

REFRENCSE

[1] Subaira.A.S, Mrs. Anitha.P, ''Efficient Classification Mechanism for Network Intrusion Detection System Based on Data Mining Techniques: a Survey'' , IEEE 2014.

[2] Nadya EL MOUSSAID, Ahmed TOUMANARI,'' Overview of Intrusion Detection Using Data-Mining and the features selection'', IEEE 2014.

[3] Subaira.A.S,P.G. Scholar, Mrs.Anitha.P,,'' Efficient Classification Mechanism for Network Intrusion Detection System Based on Data Mining Techniques :a Survey'' ,IEEE 2014.

[4] Nadya EL MOUSSAID, Ahmed TOUMANARI,'' Overview of Intrusion Detection Using Data-Mining and the features selection'', IEEE 2014.

[5] Cheung-Leung Lui ,Tak-Chung Fu, "Agent-based Network Intrusion Detection System Using Data Mining Approaches"IEEE 2005.

[6] Kalpana Jaswal, Seema Rawat,Praveen Kumar "Design Development of a for Detection using Data mining" IEEE 2015.

[7] Shengyi pan, Thaomas marris *" Developing a Hybrid Intrusion Detection System Using Data Mining for Power Syetem"*IEEE 2015. IEEE TRANSACTIONS ON SMART GRID, VOL. 6, NO. 6, NOVEMBER 2015 IEEE.

[8] Kailas Elekar, Amrit Priyadarshi M.M. Waghmare " Use of rule base data mining algorithm for Intrusion Detection" IEEE 2015. International Conference on Pervasive Computing (ICPC) 2015 IEEE

[9] Nadya EL Moussaid, Ahmed Toumanari Essi, "Overview of Intrusion Detection Using Data-Mining and the features selection*"*IEEE 2015.

[10] Ketan Sanjay Desale, Chandrakant Namdev Kumathekar, Arjun Pramod Chavan "Efficient Intrusion Detection System using Stream Data Mining Classification Technique*"2015* International Conference on Computing Communication Control and Automation.