Performance of Detection Attack using IDS Technique

[1]Baljeet, [2]Parbhat Verma,
[1]Research scholar, [2]Assistant Professor,
[1]Department of Computer Science Engineering
[1]Modern Institute of Engineering & Technology, Ambala, India

***Abstract* - With the increased use of computers and ease of access to internet, the ways to attack and deceive a system has also increased. As per WorldNet Dictionary intrusion means entering into property by force or without permission or welcome (in this case property mean computer system or network or server). For protection computer system, many methods are available; still there are many security holes. For example, firewalls cannot protect internal attacks. The essential requirement of any IDS is accuracy. The other requirements are extensibility and adaptability. The major problem with any IDS is detection of false attacks. 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,data mining, Attack,KDD , clustering.**

I. INTRODUCTION

Data mining is capable of analyzing large amount of network traffic data. It can distinguish network traffic as attack or Normal using classification algorithms. Rule base classification algorithms produce rule set by analyzing large amount of training data. And they can be tested using test [2].

Defined data mining as "The nontrivial extraction of implicit, previously unidentified, and potentially useful information from data". The major steps in Data mining are Extract, Transform, and Load (ETL). Then this data is analyzed using various techniques or algorithms for knowledge generation.

Classification is one of the commonly used supervised data mining technique. It is the process of finding a model that describes the data classes or concepts. The main function is to predict the class of objects using the model for unknown class label. The Classification model is generated using training data sets and the derived model can be presented in many forms like table, trees or rules. This is focused on rule base classification techniques such as Decision Table, JRip, OneR, PART, and ZeroR [3].

Decision Table: Decision tables (DTs) uses tabular representation for describing and analyzing situations. The decision i.e. action is taken depending upon number of conditions and their interrelationships [4].

JRip: Optimized version algorithm proposed by William W. Cohen. This algorithm will try to add every possible rule until it becomes accurate. It Optimizes rule set using discretion length [5].

OneR: It is the simplest learning algorithm for discrete attributes [6].

PART: Combines divide and conquer strategy. Incomplete C45 tree are built in each step and rule is build using best leaf [7]

ZeroR: It is the simplest classification method for simply predicting majority category (class). It relies on the target and ignores all predictors. It has no predictability power [8].

II. KDD CUP 1999 DATASET

KDD CUP 1999 is base on the intrusion detection reproduction of U.S. Air force local area networks via tcpdump [www.tcpdump.org]. The dataset consists of group contact behavior including up to 41 attributes as well as heterogeneous access patterns. The names and descriptions of the attributes are available in [9].

Attacks type	List of ATTACK
DOS	Back, land, neptune, pod, smurf, teardrop
U2R	Buffer overflow, load module, Perl, root kit
PROBES	Satan, ips weep, nmap, ports weep
R2L	Ftp_write, guesswd, imap, multihop, phf, spy, warezcl ient, warezmaster

III.CLUSTERING TECHNIQUES OR ALGORITHMS

are method in which it groups a set of objects that are similar in characteristics in one cluster. The criterion for checking the similarity is depending from algorithm to other and from clustering model to other. Hierarchical algorithm base on distance connectivity thus the connectivity models. K-means is one of the simplest clustering algorithms; it is based on centroid models that calculate the centre of each cluster by calculating the distance Euclidian among each object. former algorithms are based on

density models such as Density-based Spatial clustering of applications with noise (DB SCAN) or rdering points to identify the clustering structure (OPTICS). Graph-based models are used in algorithms where a set of nodes in a graph such that every two nodes in the set are connected by an edge can be considered from one cluster [10].

3.1 Support Vector Machine (SVM)

SVM [11][12][13][14] is a learning method for the Classification and Regression analysis of both linear and nonlinear data. It uses a hypothesis space of linear function and maps effort feature vectors into a higher dimensional space all the way through some nonlinear mapping.SVM constructs a hyper plane or set of hyper planes only the good separation is achieved by the hyper plane. [4] The hyper plane searching process in SVM is achieved by the leading margin. The related margin gives the major partition between classes. While training an SVM it create a quadratic optimization problem. In SVM the classifier is created by linear untying hyperpalne but all the linear separation cannot be solved in the original input space. SVM uses a task called kernel to solve this problem. The Kernel transforms linear problem Into nonlinear one by mapping into characteristic spaces. Radial basis function, polynomial, two layer sigmoid neural nets are the some of the kernel functions. At the time of guidance classifier, user may provide one of these functions, which selects support vectors along the surface of this function. The implementation of SVM tries to accomplish maximum separation between the classes. Intrusion detection system has two phases: training and testing. SVMs can learn a larger set of patterns and be capable to provide better classification, because the classification difficulty does not depend on the dimensionality of the feature space. SVMs also have the capability to update the training patterns dynamically whenever there is a new pattern during classification.

3.2 Genetic Algorithms

Genetic algorithms were initially introduced in the meadow of computational ecology. After that they have been bloom into various fields with promising result. Nowadays the researchers have tried to incorporate this algorithm with IDSs. The REGAL System is based on distributed genetic algorithm. REGAL is an origin learning system that learns First Order Logic multi-model concept descriptions. The learning examples are stored in relational database that are represent as relational tuples. Gonzalez and Dasgupta [15] applied a genetic algorithm, however they were examined host based IDSs, not network based. They used the algorithm only for the Meta knowledge step instead of running algorithm directly on the feature set. It uses the statistical classifiers for labelled vectors.2-bit binary encoding methodology is used for identifying the abnormality of a particular feature, ranging from normal to abnormal.Chittur [16] used a genetic algorithm with decision tree. Decision tree is used to represent the data. They used the high detection rate that reduce the false positive rate. The false positive occurrence was minimized by utilizing human input in a feedback loop.

3.3 Neural Networks

Neural Network was conventionally used to refer a network or biological neurons. [17]In IDSs neural network has been used for together anomaly and misuse intrusion detection. In anomaly intrusion detection the neural networks were modelled to identify statistically significant variations from the user's accepted behaviour also identify the typical characteristics of system users. In mistreatment 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 part into an existing system or customized expert system. This classification uses the neural network to sort the external data for suspicious events and forward them to the existing and expert system. This improves the competence 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 report of misuse attacks and identify instances that are unlike any which have been observed before by the network. It has high degree of correctness to recognize known suspicious events. Generally, it is used to learn multifaceted non linear input-output relationships.

3.4 Fuzzy Logic

Fuzzy logic is derived from fuzzy set theory it uses the rule based systems for classification. Fuzzy can be contemplation of as the application side of fuzzy set theory dealing with sound thought out real world authority values for a complex problem [17].The fuzzy data mining techniques used to remove 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 *2014* IEEE 8th Proceedings International Conference on Intelligent Systems and Control (ISCO) 277 generated. [18]The audit data and mined normal data have been compared to identify the similarity. If the similarity values are below a upper limit, an alarm raises.

3.5 Bayesian 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

3.6 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 labelled 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.

8 3.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.

IV.PROPOSED WORK

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.

Data travelling over the network is broadly classified in to two categories, normal data and anomaly. Anomaly detection is the goal of Intrusion Detection System (IDS). Different patterns can be drawn on the basis of the user's usage over the network. These patterns can be grouped together according to the similarities in them. So putting similar data into groups is called data clustering. There are many techniques for data clustering; we are using k-mean clustering which is an unsupervised learning algorithm. This technique is relies on finding cluster centers by trying to minimize a cost function of dissimilarity measure.

Step 1: initialize clustering set S to unacceptable set;
Step 2: fetch a vector d from data set;
Step 3: if S is null, build a new group centered on d, and add it to S. Go to Step 7;
Step 4: else find a cluster C j from S, which is the closest to d amongst all created clusters, that is dist(C j, d) is the smallest;
Step 5: if dist(C j, d) ≤R, add d to C j. Go to Step 7;
Step 6: else, construct a new cluster centered on d, add it to S;
Step 7: repeat (2) (3) until all vectors of statistics set are processed.
Step 8: recalculate the center of each cluster, for each cluster scanning data set from beginning, if the distance from certain vector of data set to cluster center isn't larger than R, add this vector to this cluster. Repeat it until all cluster center not changed.

V.SIMULATION RESULT

Analysis Data Set

We have performed reduction of dimensionality of the KDD Train 20 percent data set. It is an important step, not only to decrease the complexity of the training process but also to gain an insight as to which network connection features are significant for the procedure of any network intrusion detection. Having done that, these features are real information on different scales together. TCPDUMP files are converted to arff and csv format file.

Weka data mining tools [19] were used to generate naïve Bayesian and J48 classifiers with default settings and five-fold cross-validation. The results are shown in the following tables. Table 1 shows the performance of NaIve Bayesian (NB) classifier, and Table 2 shows its confusion matrix. It is wellrecognized in data mining that the following measures provide more informative evaluation of classifier performance when dealing with class-imbalanced data: recall, precision (prec.), F-measures, sensitivity, and specificity [5], which are defined as]
TP is the number of positive cases classified correctly, FN is the number of positive cases classified as negative, FP is the number of negative cases classified as positive, and TN is the number of negative cases classified correctly. In Table 2, the TP is 10298, FN is 1445 FP is 1177 and TN is 12272.

Accuracy = (TN+TP/ (TN+TP+FN+FP))

recall = TP/(TP+FN)

F-measure = (2*recall *precision) I (recall + precision),

sensitivity = TP/(TP+FN) = recall,

specificity = TN/(FP+ TN).

Table5.1: NB Output From Weka

A	b	Classified as
12272	1177	Normal
1445	10298	Anomaly

Table5.2: Bays Net Output From Weka

A	b	Classified as
13330	119	Normal
747	10996	Anomaly

Table5.3: J48Graft Output From Weka

a	b	Classified as
13425	24	Normal
25	11718	Anomaly

	Native Bays	Bays Net	J48Graft
Accuracy	89.59	96.5	99.8
Dtection Rate	87.69	93.6	97.8
False Alarm Rate	8.75	0.88	0.17
Time to Build Model	5.75	6.53	28.53

Table 5.4Comparison of NB, Bays Net, and J48 Graft

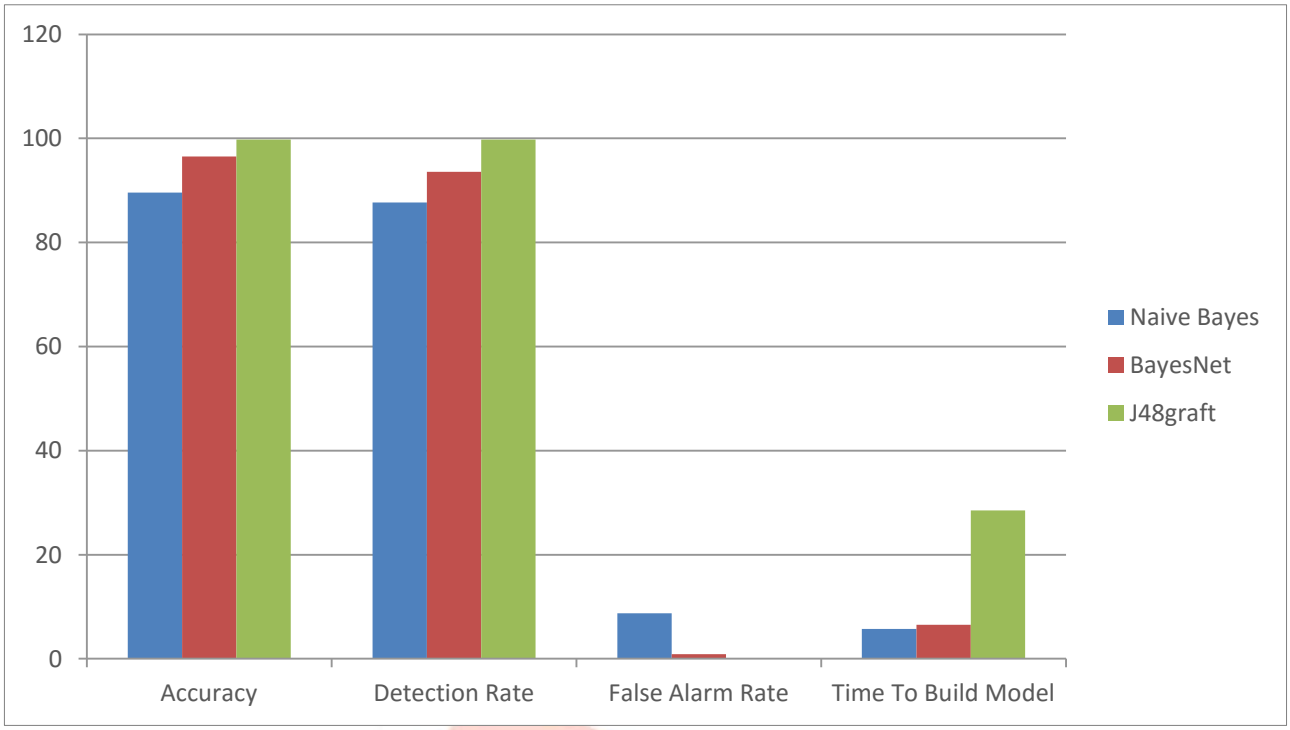

Fig5.1:Comparision between three algorithm

VI.CONCLUSION

A new training process could be simply added to the training data set without changing the weights of the existing training samples. The presentation of the k-mean clustering algorithm depends on the value of k, for k=2, the detection rate reached 96.6% rapidly and the low false optimistic rate. This could make the k mean clustering method more suitable for dynamic environments that require frequent updates of the training data.

REFERENCES

1.X Zhang, C Li , W Zheng **"Intrusion prevention system design"** , The Fourth International Conference on Computer and Information Technology, pp: 386-390, 2004
2. 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) -12015 IEEE
3.W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, "**Knowledge Discovery in Databases: An Overview,**" AI Magazine, 1992, pp. 213-228.
4 R. Kohavi,"**The Power of Decision Tables. In: 8th European Conference on Machine Learning**", pp 174-189, 1995.
5 W. Cohen"**Fast Effective Rule Induction",** In Twelfth International Conference on Machine Learning, pp 115-123, 1995.
6. R. C. Holte, "**Very Simple Classification Rules Perform Well on Most Commonly Used Datasets**," Machine Learning, pp. 63-90, 1993
7. E. Frank, H. Ian, Witten " **Generating Accurate Rule Sets Without Global Optimization"**, In: Fifteenth International Conference on Machine Learning, pp 144-151, 1998.
8. http://www.cs.**waikato.ac**.nz/ml/index.html
9.**KDD Cup 1999 Data available** at http://kdd.ics.uci.edu/databases/kddcup99/task.html, 1999.
10. Nadya EL Moussaid, Ahmed Toumanari Essi, **"Overview of Intrusion Detection Using Data-Mining and the features selection"**IEEE 2015.
11. W. Feng, Q. Zhng, G. Hu, J Xiangji Huang," **Mining network data for intrusion detection through combining SVMs with ant colony networks"Future Generation Computer Systems**,2013.
12. L. Khan, M. Awad, B. Thuraisingham**, "A new intrusion detection system using support vector** Incooperation with Columbia Univ,2001.
13. Y. Li u, X. Yu, J.X. Huang, A." **An, Combining integrated sampling with SVM ensembles forlearning from imbalanced datasets"**, Information Processing &Management (2011) 617–631.
14. S.-J. Horng , M.-Y. Su, Y.-H. Chen, " **A novel intrusion** detection system based on hierarchical clustering and support vector machines", Expert Systems with Applications 38 (2011) 306–313.
15. Dasgupta, D. and F. A. Gonzalez,"**An intelligent decision support system for intrusion detection and response",** Models and Architectures for Computer Networks Security (MMM-ACNS), 21-23 May, 2001
16.. Chittur, A., "**Model generation for an intrusion detection system using genetic algorithms",** High School Honors Thesis, Ossining High School.

[17] J. Ryan, M.-J. Lin, R. Miikkulainen,"**Intrusion detection with neural networks**", in: Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection and Task Management, 1997, pp. 92–97
18 G. J. Klir**,"Fuzzy arithmetic with requisite constraints",** Fuzzy Sets and Systems, 91:165175, methods for intrusion detection", Master's thesis, Mississippi State Univ., 1999.