ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 2, Issue 3, March 2013

Design and Implementation of Network Intrusion Detection System by using K-means clustering and Naïve Bayes

Mohan Banerjee1, Roopali Soni2
1MTech (CSE) Scholar, Department of Computer Science & Engineering, Thakral College of Technology, Bhopal,
e-mail: banerjee_mohan@rediffmail.com
2HOD & AP, Department of Computer Science & Engineering, Thakral College of Technology, Bhopal,
e-mail: rupal123_s@yahoo.com

***Abstract*— Network is the Information system(s) implemented with a collection of interconnected components. Such components may include routers, hubs, cabling, telecommunications controllers, key distribution centers, and technical control devices. Network security refers to any activities designed to protect your network. Specifically, these activities protect the usability, reliability, integrity, and safety of your network and data. Effective network security targets a variety of threats and stops them from entering or spreading on your network. Intrusion poses a serious security risk in a network environment. The ever growing new intrusion types pose a serious problem for their detection. The human labeling of the available network audit data instances is usually tedious, time consuming and expensive. In this paper, we apply one of the efficient data mining algorithms called k-means clustering via naïve bayes classification for anomaly based network intrusion detection. Experimental results on the KDD cup'99 data set show the novelty of our approach in detecting network intrusion. It is observed that the proposed technique performs better in terms of Detection rate when applied to KDD'99 data sets compared to a naïve bayes based approach.**

***Keywords*— Network Intrusion Detection, K-Means Clustering, Naïve Bayesian Classification, Detection Rate and False Positive Rates.**

I. INTRODUCTION

The networks are computer networks, both public and private, that are used every day to conduct transactions and communications among businesses, government agencies and individuals. The networks are comprised of "nodes", which are "client" terminals (individual user PCs), and one or more "servers" and/or "host" computers. They are linked by communication systems, some of which might be private, such as within a company and others which might be open to public access. The obvious example of a network system that is open to public access is the Internet, but many private networks also utilize publicly-accessible communications. Today, most companies' host computers can be accessed by their employees whether in their offices over a private communications network, or from their homes or hotel rooms while on the road through normal telephone lines. Network security involves all activities that organizations, enterprises, and institutions undertake to protect the value and ongoing usability of assets and the integrity and continuity of operations. An effective network security strategy requires identifying threats and then choosing the most effective set of tools to combat them. With the tremendous growth of network-based services and sensitive information on networks, network security is becoming more and more importance than ever before. Intrusion detection techniques are the last line of defenses against computer attacks behind secure network architecture design, firewalls, and personal screening. Despite the plethora of intrusion prevention techniques available, attacks against computer systems are still successful. Thus, intrusion detection systems (IDSs) play a vital role in network security. Symantec in a recent report[1] uncovered that the number of fishing attacks targeted at stealing confidential information such as credit card numbers, passwords, and other financial information are on the rise, going from 9 million attacks in June2004 to over 33 millions in less than a year.

One solution to this is the use of network intrusion detection systems (NIDS), which detect attacks by observing various network activities. It is therefore crucial that such systems are accurate in identifying attacks, quick to train and generate as few false positives as possible. This paper presents the scope and status of our research in anomaly detection. This paper gives a comparative study of k-means clustering via naïve bayes classification and naïve bayes classification for identifying novel network intrusion detections. We present experimental results on KDDCup'99 data set. Experimental results have demonstrated that our k-means clustering via naïve bayes classifier model is much more efficient in the detection of network intrusions, compared to the naïve bayes classification based classification techniques. Section 2 describes IDS in general. Section 3 presents an overview of frequently occurring network attacks, and section 4 discusses related research done so far. Section 5 describes our proposed method and section 6 presents the experimental results. Finally, section 7 provides the concluding remarks and future scope of the work.

II. DATASET DESCRIPTION

The network connection records used in the experiment were the KDD CUP 1999 Dataset from MIT's Lincoln Lab, which has

been developed for IDS evaluation by DARPA. It consists of approximately 494020 data instances, each of which is a vector of extracted feature values from a connection record obtained from the raw network data gathered during the simulated intrusions and a vector consisting of 42 various quantitative and qualitative features.
Four different categories of attack patterns are included:

a. Denial of Services attack (DOS). Examples are Apache2, Land, Mail bomb, SYN Flood, Ping of death, Process table, Smurf and Syslogd
b. User to Super user or Root Attacks (U2R). Examples are Eject, Ffbconfig, Fdformant, loadmodule, Perl, Ps and Xterm.
c. Remote to User Attack (R2L). Examples are Dictionary, Ftp_write, Gest, Imap, Named, Phf, Sendmail , Xlock and Xsnoop
d. Probing: probing is a class of attacks in which an attack scans a network of computers to gather information or find known vulnerabilities.

III. RELATED WORK

In [10], Axellson wrote a well-known paper that uses the Bayesian rule of conditional probability to point out that implication of the base-rate fallacy for intrusion detection. In [11], a behaviour model is introduced that uses Bayesian techniques to obtain model parameters with maximal a-posteriori probabilities.

IDDM (Intrusion Detection using Data Mining Technique) [5] is a real-time NIDS for misuse and anomaly detection. It applies association rules, meta rules, and characteristic rules. It employs data mining to produce description of network data and uses this information for deviation analysis.

ADAM (Audit Data Analysis and Mining) [4] is an intrusion detector built to detect intrusions using data mining techniques. It first absorbs training data known to be free of attacks. Next, it uses an algorithm to group attacks, unknown behaviour, and false alarms. ADAM has several useful capabilities, namely; Classifying an item as a known attack. Classifying an item as a normal event. Classifying an item as an unknown attack. Match audit trial data to the rules it gives rise to.

MADAM ID (Mining Audit Data for Automated Models for Intrusion Detection) [6] is one of the best known data mining projects in intrusion detection. It is an off-line IDS to produce anomaly and misuse intrusion detection models. Association rules and frequent episodes are applied in MADAM ID to replace hand-coded intrusion patterns and profiles with the learned rules.

In [7], the authors propose a method of intrusion detection using an evolving fuzzy neural network. This type of learning algorithm combines artificial neural network (ANN) and fuzzy Inference systems (FIS), as well as evolutionary algorithms. They create an algorithm that uses fuzzy rules and allow new neurons to be created in order to accomplish this. They use Snort to gather data for training the algorithm and then compare their technique with that of an augmented neural network.

In [8], a statistical neural network classifier for anomaly detection is developed, which can identify UDP flood attacks. Comparing different neural network classifiers, the back propagation neural network (BPN) has shown to be more efficient in developing IDS [9]. In [9], the author uses the back propagation method by Sample Query and Attribute Query for the Intrusion Detection, whereby analysing and identifying the most important components of training data. It could reduce processing time, storage requirement, etc.

IV. THE PROPOSED ALGORITHM

Proposed Algorithm:

Input: D is Kddcup database

Output: Intrusion detection Model

Learning Algorithm

Step 1: Define the number of cluster on basic of similarity.

(K>=n)

Where K = no. of cluster

n = no. of data sample.

Step 2: Clustering on the bases of similarity

Step 3: Calculating the centroids of clustering

$$J_m= \Sigma_{(i=1)}^{N} \Sigma_{(J=1)}^{C} u_{(ij} \| X_i - C_j \|)^m 2$$

Where, Xt is a vector representing the t-th data point in the cluster Ci and ci is the geometric centroid of the cluster Ci. Finally, this algorithm aims at minimizing an objective function, in this case a squared error-function, where ||Xt - Ci||2 is a chosen distance measurement between data point xt and the cluster centre Ci [41]

Step 4: Calculate the prior probability using

P (Ki) =Si ⎪ S

Ki=No. of Cluster label in Kddcup database.

Si=No. of training Sample for the each cluster.

S=Total number of Sample.

Step 5: Calculate the posterior probability

Here we are calculate the posterior probability [41]

based on the Kddcup dataset

D=P (H ⎪ X)

P (H□ X)= P(⊢ X⊣ |H)P(H)/P(X)

H=Hypothesis

X=Kddcup data sample.

Naive Bayes it makes the assumption that all the input attributes are independent,such as one attribute doesn't affect the other in deciding whether or not a condition in the database.

Step 6: For each attribute Ai the number of occurrences of each attribute value X can be counted to determine P (Ai).

P (A ⎪ C) =P (Ai) ⎪ P(Ki)

Here

Ai=No. of attribute are used in the D

Ki=No. of clusters in the D

Where i=0, 1,2,3,4,

Step 7: Classify all the training examples using these prior and conditional probabilities,

P (Ci□ X)= P(├ X┤ |Ci)P(Ci)/P(X)

Then P (Ki) would be the probability that the how many clusters in the Dataset, and P (X|K) is the probability that the given condition (input) in the dataset, given that the condition in the dataset. P(X) is just the probability of a condition appearing in database.

Step 8: Calculate the maximum posterior probability of dataset.

Step 9: Again classify all training examples in D using updated probability values.

V. EXPERIMENT AND RESULTS

For our experiments we are using KDD CUP 99 dataset. KDD CUP 1999 contains 41 fields as an attributes and 42nd field as a label. In our algorithm we have taken selected features. The 42nd field can be generalized as Normal, DoS, Probing, U2R, and R2L. The description of KDD CUP 99 used for our method shown in table 1. The performances of each method are measured according to the Accuracy, Detection Rate and False Positive Rate using the following expressions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Detection \quad Rate = \frac{TP}{TP + FP}$$

$$False \quad Alarm = \frac{FP}{FP + TN}$$

Where,

FN is False Negative,TN is True Negative,

TP is True Positive, andFP is False Positive

The detection rate is the number of attacks detected by the system divided by the number of attacks in the data set. The false positive rate is the number of normal connections that are misclassified as attacks divided by the number of normal connections in the data set.

A "Confusion Matrix" is sometimes used to represent the result of , as shown in Table .The Advantage of using this matrix is that it not only tells us how many got misclassified but also what misclassifications occurred. For our model we get the confusion matrix shown in Table 2, Table3, Table 4 and Table 5.

Attack Types	Number of Records
Normal	97277
Denial of Service	391458
Remote to User	1126
User to Root	52
Probing	4107
Total Examples	494020

Table 1: shows the number of examples in 10% training and testing data of KDD99 dataset.

A. Confusion Matrix for Naïve Bayesian Classifier

In this work research naïve Bayes and k-Means with NB algorithms are used to classify attacks from kddcp99 data [45] sets. Confusion matrix obtained from above methods is as follows.

Actual	Predicted Normal	Predicted DOS	Predicted Probing	Predicted U2R	Predicted R2L
Normal	71586	17859	5876	1813	144
DOS	2282	388676	485	15	0
Probing	1011	2628	195	273	0
U2R	16	3	0	86	0
R2L	678	1	9	304	81

Table 2.Confusing Matrix for Naïve Bayes

Above matrix horizontal shown the predicted attack in the kddcup1999 data set and vertically shown the actual class in the Kddcup data set. Here total number of normal data, Dos attack, R2L, U2R and probing is 60593,229853,4595,11822 and 4166.Using all the above three parameters accuracy, false positive rate and precision rate are computed.

B. Result for Naïve Bayes:

This matrix had shown the precision rate, accuracy and false positive rate using Naïve Bayesian classifier.

Attack	Detection Rate	False Alarm	Accuracy
Normal	73.59	6.19	93.95
Dos	99.29	3.72	95.19
Probing	4.75	0.84	97.82
U2R	6.65	0.00	99.48
R2L	7.55	0.21	99.75

Table 3 Precision rate, accuracy and false positive rate compute by Naïve Bayesian classifier

C. Confusion matrix for with K-means algorithms Naïve Bayesian

The following result Table 4 is concluded by use of hybrid algorithms. This matrix shown, whether the attack performed by Naïve Bayesian with K-means is classified correctly or

incorrectly. This confusion matrix is produce by MATLAB Tool 7.9.

Actual	Predicted Normal	Predicted DOS	Predicted Probing	Predicted U2R	Predicted R2L
Normal	83694	2584	1084	18	445
DOS	7289	369038	383	2	1
Probing	5382	19825	2405	0	9
U2R	343	3	232	73	16
R2L	570	8	0	12	602

Table 4 Resulting Confusion Matrix of KNMB

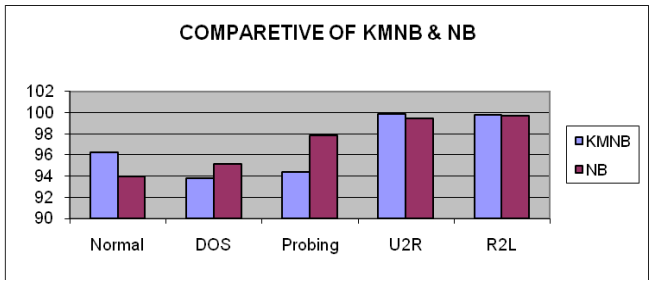
D. Result for Naïve Bayes with K-means Algorithms:

This matrix is shown the result for hybrid algorithms.

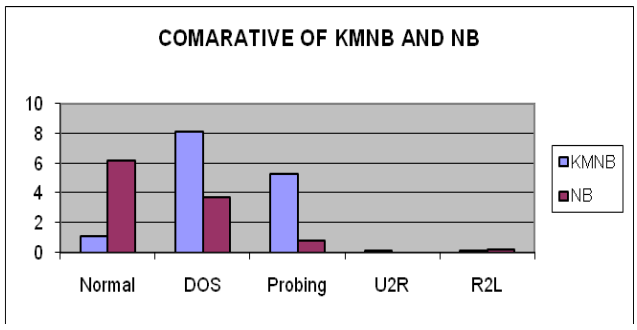
Attack	Detection Rate	False Alarm	Accuracy
Normal	95.30	1.10	96.26
DOS	97.96	8.13	93.81
Probing	8.71	5.27	94.42
U2R	0.00	0.13	99.86
R2L	50.50	0.13	99.77

Table 5: Precision rate, accuracy and false positive rate compute by Naïve Bayesian classifier with k-means algorithms

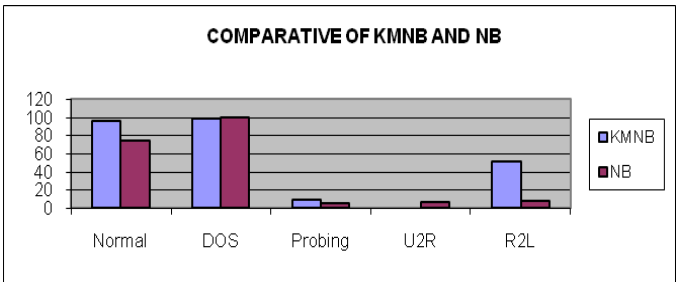
1,2 and 3 gives a respectively comparative of Accuracy , False Alarm and Detection Rate performance of experimental result of combine approach k-means clustering and naïve bayes classification .



Graph 1. Accuracy Rate of KMNB



Graph 2. False Alarm of KMNB



Graph 3. Detection Rate of KMNB

In this overall measurement we have discussed about the experimental results for a single classifier Naïve Bayes and KMNB are summarized in representing measurement in terms of accuracy, precision rate and false alarm on the data set. From the Naïve Bayes classifier has produced a high accuracy and precision rate in some attack but with high false alarm rates. In hybrid algorithms, K-Means with Naïve Bayesian classifier (KMNB) produces improved accuracy and precision rate with low false alarm rates. The false alarm for the single Naïve Bayes classifier increased up to 8.4% with moderate accuracy and precision rates which 90%. The clustering techniques used as a pre-classification component for grouping similar data by classes in the earlier stage helps KMNB produces a improved result compared to Naïve Bayes classifier. The data which misclassified during the first stage was classified accordingly in the second stage; hence making KMNB outperforms Naïve Bayes classifier in term of false alarm.

VII. Conclusion and Future Work

In this paper we have used two learning algorithms of data mining i.e. K-means and Naive Bayes classifier. K-means is a clustering algorithm, which work to provide grouping to data sample on the basis of their similarities and dissimilarities. Naive Bayes classifier is classification algorithm which correctly classifies the intrusion/attack. The combination of these two algorithms used in order to improve accuracy, precision rate and reduce the false positive rate. In this paper, we apply one of the efficient data mining algorithms called k-means clustering via naïve bayes classification for anomaly based network intrusion detection. Experimental results on the KDD cup'99 data set show the novelty of our approach in detecting network intrusion. It is observed that the proposed technique performs better in terms of Detection rate when applied to KDD'99 data sets compared to a naïve bayes based approach.

Its future work is to increase accuracy, Detection rate of DOS attack and reduce the false positive rate.

REFERENCES

[1] "Symantec-Internet Security threat report highlights (Symantec.com)", http://www.prdomain.com/companies/Symantec/newreleases/Symantec_internet_205032.htm

[2] R.Durst, T.champion, B.witten, E.Miller, and L.Spagnuolo, "Testing and valuating computer intrusion detection system" communications of ACM, Vol.42, no.7, pp 53-61, 1999. \

[3] A.Sung & S.Mukkamala, "Identifying important features for intrusion detection using SVM and neural networks, " in symposium on application and the Internet, pp 209-216, 2003.

[4] D.Barbara, J.Couto, S.Jajodia, and N.Wu, "ADAM: A test bed for exploring the use of data mining in intrusion detection", SIGMOD, vol30, no.4, pp 15-24, 2001.

[5] Tomas Abraham, "IDDM: INTRUSION Detection using Data Mining Techniques", Technical report DSTO electronics and surveillance research laboratory, Salisbury, Australia, May2001.

[6] Wenke Lee and Salvatore J.Stolfo, "A Framework for constructing features and models for intrusion detection systems", ACM transactions on Information and system security (TISSEC), vol.3, Issue 4, Nov 2000.

[7] S.chavan, K.Shah, N.Dave, S.Mukherjee, A.Abraham, and S.Sanyal, Adaptive neuro-fuzzy Intrusion detection syatems", ITCC, Vol 1, 2004

[8] Z. Zhang, J. Li, C.N. Manikapoulos, J.Jorgenson, J.ucles, "HIDE: A hierarchical network intrusion detection system using statistical pre-processing and neural network classification ", IEEE workshop proceedings on Information assurance and security, 2001, pp.85-90.

[9] Roy-I Chang, Liang-Bin Lai, et al, "Intrusion detection by back propagation network with sample query and attribute query ", International Journal of computational Intelligence Research, Vol.3, no.1, 2007, pp 6-10.

[10] S. Axelsson, "The base rate fallacy and its implications for the difficulty of Intrusion detection", Proc. Of 6th.ACM conference on computer and communication security 1999.

[11] R.Puttini, Z.marrakchi, and L. Me, "Bayesian classification model for Real time intrusion detection", Proc. of 22nd. International workshop on Bayesian inference and maximum entropy methods in science and engineering, 2002.

[12] MacQueen, .Some methods for classification and analysis of multivariate observations in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1967, pp. 281.297.

[13] KDD99. KDD99 cup dataset http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html,1999.

[14] P.Jenson, "Bayesian networks and decision graphs", Springer, New-york, USA, 2001.

[15] J.Pearl, "Probabilistic reasoning in intelligent system", Networks of plausible inference, Morgan Kaufmann 1997.

[16] S.J.Russel, and Norvig, "Artificial Intelligence: A modern approach "(International edition), Pearson US imports & PHIPES, Nov 2002.

[17] Mrutyunjaya Panda and Manas Ranjan Patra , "NETWORK INTRUSION DETECTION USING NAÏVE BAYES" IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.12, December 2007.