An Improved Network Intrusion Detection Technique based on k-Means Clustering via Naïve Bayes Classification

Sanjay Kumar Sharma[1], Pankaj Pandey[1], Susheel Kumar Tiwari[2] and Mahendra Singh Sisodia[1], *Member, IEEE*
[1]Assistant Professor, Department of Computer Science & Engineering
[2]Assistant Professor, Department of Computer Science & Engineering
[1]Oriental Institute of Science & Technology, Bhopal, India
[2]Millennium Institute of Technology & Science, Bhopal, India
[1]{sanjaysharma, pankajpandey, mahendrasisodia}@oriental.ac.in
[2]sushiltiwari24@yahoo.co.in

***Abstract*— As network attacks have increased in number and severity over the past few years, intrusion detection system (IDS) is increasingly becoming a critical component to secure the network. Due to large volumes of security audit data as well as complex and dynamic properties of intrusion behaviors, optimizing performance of IDS becomes an important open problem that is receiving more and more attention from the research community. 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, ROC graph, Detection Rate and False Positive Rates.**

I. INTRODUCTION

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. INTRUSION DETECTION

An Intrusion Detection System (IDS) inspects the activities in a system for suspicious behaviour or patterns that may indicate system attack or misuse. There are two main categories of intrusion detection techniques; Anomaly detection [2] and Misuse detection. The former analyses the information gathered and compares it to a defined baseline of what is seen as "normal" service behaviour, so it has the ability to learn how to detect network attacks that are currently unknown. Misuse Detection is based on signatures for known attacks, so it is only as good as the database of attack signatures that it uses for comparison. Misuse detection has low false positive rate, but cannot detect novel attacks. However, anomaly detection can detect unknown attacks, but has high false positive rate.

In this paper, we review the performance of k-means clustering and naïve classifier when trained to identify

ISBN: 978-81-909042-2-3

signatures of specific attacks. These attacks are discussed in more detail in the following section.

III. NETWORKING ATTACK

The simulated attacks were classified, according to the actions and goals of the attacker. Each attack type falls into one of the following four main categories [3]:

Denials-of Service (DoS) attacks have the goal of limiting or denying services provided to the user, computer or network. A common tactic is to severely overload the targeted system. (e.g. apache, smurf, Neptune, Ping of death, back, mailbomb, udpstorm, SYNflood, etc.).

Probing or Surveillance attacks have the goal of gaining knowledge of the existence or configuration of a computer system or network. Port Scans or sweeping of a given IP-address range typically fall in this category. (e.g. saint, portsweep, mscan, nmap, etc.).

User-to-Root (U2R) attacks have the goal of gaining root or super-user access on a particular computer or system on which the attacker previously had user level access. These are attempts by a non-privileged user to gain administrative privileges (e.g. Perl, xterm, etc.).

Remote-to-Local(R2L) attack is an attack in which a user sends packets to a machine over the internet, which the user does not have access to in order to expose the machine vulnerabilities and exploit privileges which a local user would have on the computer (e.g. xclock, dictionary, guest_password, phf, sendmail, xsnoop, etc.).

IV. RELATED WORK

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.

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.

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.

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.

V. THE PROPOSED MODEL FOR NIDS

Our NIDS approach deploys the K-mean clustering algorithm [12] make cluster with normal and anomalous traffic in the training dataset. The resulting cluster centroids are then used for fast anomaly detection in new monitoring data. The raw data and the extracted features that serve as input for the data mining algorithm. Finally, we show how the patterns can be used for classification by naïve bayes and outlier detection shown in figure 1.

A. Dataset Description

The data set used was the KDD Cup 1999 Data [13], which contained a wide variety of intrusions simulated in a military network environment. It consisted of approximately 4,900,000 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. The simulated attacks fell in one of the following four categories: DOS-Denial of Service (e.g. a syn flood), R2L- Unauthorized access from a remote machine (e.g. password guessing), U2R-Unauthorized access to superuser or root functions (e.g. a buffer overflow attack), Probing-surveillance and other probing for vulnerabilities (e.g. port scanning).

B. Feature Selection

The feature selection included the basic features of an individual TCP connection such as its duration, protocol type, number of bytes transferred, and the flag indicating the normal or error status of the connection. Other features of an individual connection obtained using some domain knowledge, and included the number of file creation operation, number of failed login attempts. In total, there were 41 features, with most of them taking on continuous values.

ISBN: 978-81-909042-2-3 ©2012 IEEE

A. Normalization

Since our algorithm is designed to be general, it must be able to create clusters given a dataset from an arbitrary distribution. A problem with typical data is that different features are on different scales. This cause bias toward some features over other features. To solve this problem, we convert the data instances to a standard form based on the training dataset's distribution. That is, we make the assumption that the training dataset accurately reflects the range and deviation of feature values of the entire distribution. Then, we can normalize all data instances to a fixed range of our choosing, and hard code the cluster width based on this fixed range.

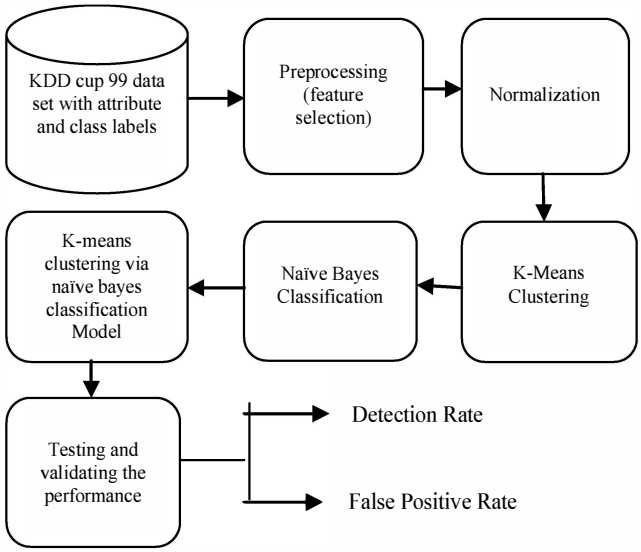

Figure 1: k-means clustering via naïve bayes classification model for NIDS

B. K-Means Clustering

The K-means clustering [12] is a clustering analysis algorithm that groups objects based on their feature values into K disjoint clusters. Objects that are classified into the same cluster have similar feature values. K is a positive integer number specifying the number of clusters, and has to be given in advance. Here are the four steps of the K-means clustering algorithm:

1) Define the number of clusters K.
2) Initialize the K cluster centroids. This can be done by arbitrarily dividing all objects into K clusters, computing their centroids, and verifying that all centroids are different from each other. Alternatively, the centroids can be initialized to K arbitrarily chosen, different objects.
3) Iterate over all objects and compute the distances to the centroids of all clusters. Assign each object to the cluster with the nearest centroid.
4) Recalculate the centroids of both modified clusters.
5) Repeat step 3 until the centroids do not change any more.

A distance function is required in order to compute the distance (i.e. similarity) between two objects. The most commonly used distance function is the Euclidean one which is defined as:

$$d(x,y) = \sqrt{\sum_{i=1}^{m} (x^i - y^i)^2}$$

where $x = (x_1,...x_m)$and $y = (y_1,...y_m)$are two input vectors with m quantitative features. In the Euclidean distance function, all features contribute equally to the function value. However, since different features are usually measured with different metrics or at different scales, they must be normalized before applying the distance function.

C. Naïve Bayesian Classification

The Naïve Bayes method is based on the work of Thomas Bayes (1702-1761). In Bayesian classification, we have a hypothesis that the given data belongs to a particular class. We then calculate the probability for the hypothesis to be true. This is among the most practical approaches for certain types of problems. The approach requires only one scan of the whole data. Also, if at some stage there are additional training data, then each training example can incrementally increase/decrease the probability that a hypothesis is correct. Thus, a Bayesian network is used to model a domain containing uncertainty [14, 15].

The naïve Bayes model is a heavily simplified Bayesian probability model [16]. In this model, consider the probability of an end result given several related evidence variables. The probability of end result is encoded in the model along with the probability of the evidence variables occurring given that the end result occurs. The probability of an evidence variable given that the end result occurs is assumed to be independent of the probability of other evidence variables given that end results occur. Now we will consider the alarm example using a naïve Bayes classifier. Assume that we have a set of examples that monitor some attributes such as whether it is raining, whether an earthquake has occurred etc. Lets assume that we also know, using the monitor, about the behaviour of the alarm under these conditions. In addition, having knowledge of these attributes, we record whether or not a theft actually occurred. We will consider the category of whether a theft occurred or not as the class for the naïve Bayes classifier. This is the knowledge that we are interested in. The other attributes will be considered as knowledge that may give us evidence that the theft has occurred.

In the training phase, the naïve bayes algorithm calculates the probabilities of a theft given a particular attribute and then stores this probability. This is repeated for each attribute, and the amount of time taken to calculate the relevant probabilities

ISBN: 978-81-909042-2-3 ©2012 IEEE

for each attribute. In the testing phase, the amount of time taken to calculate the probability of the given class for each example in the worst case is proportional to n, the number of attributes. However, in worst case, the time taken for testing phase is same as that for the training phase.

VI. 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, and

FP 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	Training Examples	Testing Examples
Normal	97277	60592
Denial of Service	391458	237594
Remote to User	1126	8606
User to Root	52	70
Probing	4107	4166
Total Examples	494020	311029

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

Actual	Predicted Normal	Predicted DoS	Predicted Probe	Predicted U2R	Predicted R2L	Accuracy (%)
Normal	8909	8	138	570	102	91.6
DoS	444	36921	16	1757	8	94.3
Probe	0	0	410	0	1	99.8
U2R	0	0	0	4	1	80.0
R2L	27	0	3	9	74	65.5

Table 2: Confusion Matrix for naïve bayes classifier using training data set.

Actual	Predicted Normal	Predicted DoS	Predicted Probe	Predicted U2R	Predicted R2L	Accuracy (%)
Normal	9687	3	23	5	9	99.6
DoS	3	33936	0	0	207	99.5
Probe	0	0	410	0	0	100
U2R	1	0	0	2	2	40.0
R2L	35	2	3	4	69	61.6

Table 3: Confusion Matrix for K-Means clustering by naïve Bayes classification using training data set

Actual	Predicted Normal	Predicted DoS	Predicted Probe	Predicted U2R	Predicted R2L	Accuracy (%)
Normal	7875	14	131	1664	43	81.0
DoS	6431	32298	417	0	0	82.5
Probe	6	12	393	0	0	95.6
U2R	1	0	0	4	0	80.0
R2L	10	00	1	0	102	90.3

Table 4: Confusion Matrix for naïve bayes classifier using testing data set.

ISBN: 978-81-909042-2-3 ©2012 IEEE

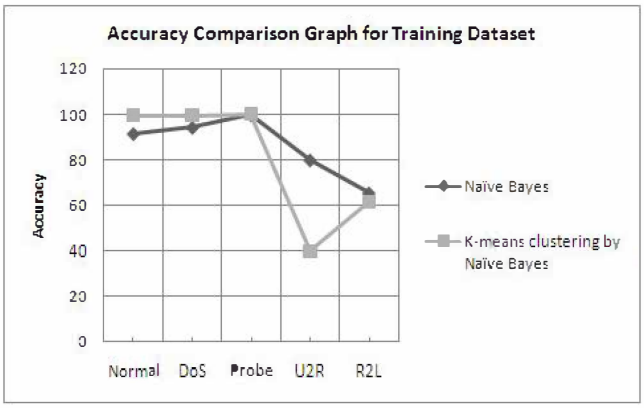

Figure 2: Accuracy comparison graph by using training data set

Actual	Predicted Normal	Predicted DoS	Predicted Probe	Predicted U2R	Predicted R2L	Accuracy (%)
Normal	9678	9	3	35	2	99.5
DoS	134	38984	27	0	1	99.6
Probe	0	3	404	4	0	98.3
U2R	1	0	0	4	0	80.0
R2L	4	12	0	3	94	98.3

Table 5: Confusion Matrix for K-Means Clustering via Naïve Bayesian classification using testing data set.

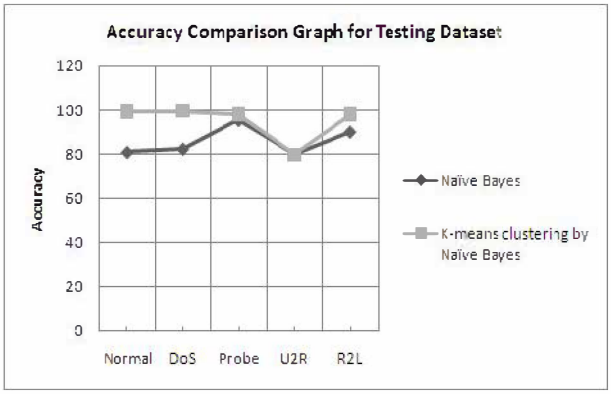

Figure 3: Accuracy comparison graph by using testing data set.

Figure 2 and Figure 3 show the comparison of accuracy for our method and naïve bayes classification. In [17] which uses Naïve Bayesian Classification shows that the detection rate in detecting intrusion is 95% . However, in our case, the detection rate is 99%, with an error rate of 4%. However, in comparison to Naïve Bayesian Classification, our approach generates more false positives.

VII. Conclusion and Future Work

In this paper, we have proposed a model of NIDS based on K-Means Clustering via Naïve Bayes algorithm. The model builds the patterns of the network services over data sets labelled by the services. With the built patterns, the model detects attacks in the datasets using the k-means clustering via naïve Bayes Classifier algorithm. Compared to the Naïve based approach, our approach achieve higher detection rate. However, it generates somewhat more false positives

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.

ISBN: 978-81-909042-2-3 ©2012 IEEE

Mr. Mahendra Singh Sisodia received M. Tech in Computer Science & Engineering and currently working as Assistant Professor in Oriental Institute of Science & Technology, Bhopal (M.P.) India. He has 6 years teaching experience and he has published two (International Journal), five (International Conference) and two (National Conference). His research area in Network Security , Data Mining, Web Mining, Algorithm Optimization, Machine Learning etc.

Mr. Sanjay Kumar Sharma received M. Tech in Computer Science & Engineering and currently working as Assistant Professor in Oriental Institute of Science & Technology, Bhopal (M.P.) India. He has 8 years teaching experience and he has published two (International Journal), two (International Conference) and three (National Conference). His research area in Ad hoc Network, Network Security

Mr. Pankaj Pandey received B.E. in Computer Science & Engineering and currently working as Assistant Professor in Oriental Institute of Science & Technology, Bhopal (M.P.) India. He has 6 years teaching experience and he has published one (International Journal), two (International Conference) and two (National Conference). His research area in Network Security , Data Mining, Software Engineering, Machine Learning etc.

Susheel Kumar Tiwari received M. Tech in Information Technology and currently working as Assistant Professor in Millennium Institute of Technology & Science, Bhopal (M.P.) India. He has 6 years teaching experience His research area in Network Security , Data Mining.

ISBN: 978-81-909042-2-3 ©2012 IEEE