Intrusion Detection Systemby Improved Preprocessing Methods and Naïve Bayes Classifier using NSL-KDD 99 Dataset

Datta H.Deshmukh
Department Of Computer Engineering
Ramrao Adik Institute Of Technology
Navimumbai,INDIA
deshmukh.datta7@gmail.com

Tushar Ghorpade
Department Of Computer Engineering
Ramrao Adik Institute Of Technology
Navimumbai,INDIA
tushar.ghorpade@gmail.com

Puja Padiya
Department Of Computer Engineering
Ramrao Adik Institute Of Technology
Navimumbai,INDIA
puja .padiya@gmail.com

Abstract—**Today Network is one of the very important parts of life and a lot of essential activities are performed using network. Network security plays critical role in real life situations. This paper presents a Data Mining method in which various preprocessing methods are involved such as Normalization, Discretization and Feature selection. With the help of these methods the data is preprocessed and required features are selected. Here Naïve Bayes classifier is used in supervised learning method which classifies various network events for the KDD cup'99 Dataset. This dataset is the most commonly used dataset for Intrusion Detection.**

***Keywords*–Correlation Based Feature Selection,Naive Bayes,Cross validation,Normalization,Discretization,Knowledge Discovery in Databases.**

I. INTRODUCTION

As the computer network usage is increased a lot of vital applications running on it also increased therefore network security is very important issue [1].As the network grows various attacks on it also increased at constant rate. Intrusion detection is a process of analyzing and monitoring various activities of network in order to detect signs of security problem .i.e. IDS is one way of dealing with suspicious activity within a network and if any malicious activity is found then it produces a report to the management station [2]. IDS can be categorized it misuse detection and anomaly detection. In which anomaly detection is based on behavior in which any action that significantly deviate from normal behavior comes under intrusion [3].

Data mining techniques are used to explore and analyze large dataset and find useful patterns [4].Classification is the category that consists of identification of class labels of records that are typically described by set of features in dataset [5].

The aim of this paper is to develop the Intrusion Detection model which classify network event as normal or attack event for which supervised learning is used. This is the machine learning technique in that it is told to which class each training tuple belongs. In this data analysis task is classification, where a classifier model is constructed to predict categorical labels. Classification is the identification of the category labels of instances that are typically described by a set of features in a dataset [6].

II. RELATED WORK

This paperpresent a literature review of few areas that covers span of research .The data set is publicly available for researchers through the website.

Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani [7] conducted a statistical analysis on this data set; they found some important issues which highly affects the performance of evaluated systems, and results in a very poor evaluation of anomaly detection approaches. To solve these issues, they have proposed a new data set, NSL-KDD [8] which has the following advantages over the original KDD data set:

1. It does not include redundant records in the train set, so the classifiers will not be biased towards more frequent records.

2. There is no duplicate records in the proposed test sets; therefore, the performance of the learners are not biased by the methods which have better detection rates on the frequent records.

3. The number of selected records from each difficulty level group is inversely proportional to the percentage of records in the original KDD data set. As a result, the classification rates of distinct machine learning methods vary in a wider range, which makes it more efficient to have an accurate evaluation of different learning techniques.

4. The number of records in the train and test sets are reasonable, which makes it affordable to run the experiments on the complete set without the need to randomly select a small portion. Consequently, evaluation results of different research works will be consistent and comparable.

Adetunmbi A.Olusola. Adeola S.Oladele. And Daramola O.Abosede [9] presented the relevance of each feature in KDD '99 intrusion detection dataset to the detection of each class. Rough set degree of dependency and dependency ratio of each class were employed to determine the most discriminating features for each class. Selecting the right features is

challenging, but it must be performed to reduce the number of features for the sake of efficient processing speed and to remove the irrelevant, redundant and noisy data for the sake of predictive accuracy.A multiclass classifier G needs to map the given inputs with A features into C classes on a dataset D, which consists of {E1,E2,…, Ei, …,Et} instances.

H. GüneşKayacık, A. NurZincir-Heywood, Malcolm I. Heywood has given a feature relevance analysis. It is performed on KDD 99 training set, which is widely used by machine learning researchers. Feature relevance is expressed in terms of information gain, which gets higher as the feature gets more discriminative. In order to get feature relevance measure for all classes in training set, information gain is calculated on binary classification, for each feature resulting in a separate information gain per class. Recent research employed decision trees, artificial neural networks and a probabilistic classifier and reported, in terms of detection and false alarm rates, that user to root and remote to local attacks are very difficult to classify. The contribution of this work is that it analyses the involvement of each feature to classification.

Panda and Patra [10] applied a classifier based on a naïve Bayes method to the intrusion detect problem. They carried out their experiments using KDD'99 data set with four attack classes and compared the performance of their model in terms of error rate and cost with a model based on neural networks using K-means clustering. Based on their analysis, their approach achieve higher detection rate, less time consuming and has low cost factor while it generates more false positives.

III. DATASET

The KDD Cup 99 dataset, which derived from the DARPA IDS evaluation dataset, was used for the KDD Cup 99 Competition (KDD Cup 99 Dataset, 2009). The complete dataset has almost 5 million input patterns and each record represents a TCP/IP connection that is composed of 41 features that are both qualitative and quantitative in nature The dataset used in our study is a smaller subset (10% of the original training set), that contains 494,021 instances and it was already employed as the training set in the competition. For the test set, we used the original KDD Cup 99 dataset containing 331,029 patterns. It consists of four attack types are shown.

i. Denial of Service (DoS): attacks, where an attacker makes some computing or memory resource too busy or too full to handle legitimate requests, thus denying legitimate users access to a machine.

ii. Probe attacks: where an attacker scans a network to gather information or find known vulnerabilities.

iii. Remote-to-Local (R2L) attacks: where an attacker sends packets to a machine over a network, then exploits machines vulnerability to illegally gain local access as a user.

iv. User-to-Root (U2R) attacks**:** where an attacker startsout with access to a normal user account on the system and is able to exploit vulnerability to gain root access to the system.

IV. THE PROPOSED METHOD

1. Split dataset into two separate sets: training set and test set.
2. Perform pre-processing.
3. Build a model by learning from the training set. Refine the model by cross validation on the training set.
4. Evaluate the trained model(s) on the test sets.

Detailed Steps to be performed for Intrusion Detection Framework are as follows:

1. Initially it consists of the training set and test set.

2. Next the Min-max normalization subtracts the minimum value of an attribute from each value of the attribute and then divides the difference by the range of the attribute. These new values are multiplied by the new range of the attribute and finally added to the new minimum value of the attribute. These operations transform the data into a new range, generally [0, 1].

3. Although the NB classifier model is based on discrete features, the KDD'99 dataset mainly consists of continuous features, which need to be first converted to discrete features.

4.Final pre-processing task based on proposed model, includes a feature selection model based on the filter methods: CFS i.e. Correlation-Based Feature Selection.

5. These approaches are leading filter-based feature selection methods that provided good results in the naïve Bayes method on the KDD'99 dataset as shown in Proposed System figure 4.1.

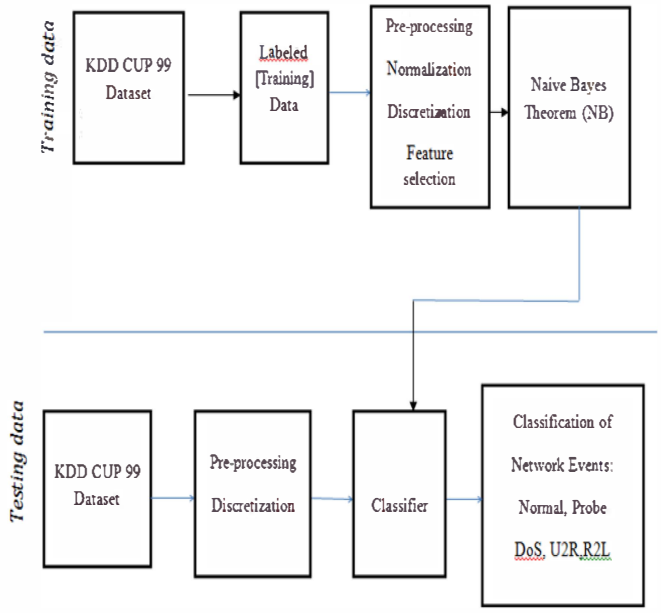

FIGURE 4.1 PROPOSED SYSTEM

4.1 Data Pre-Processing

Following steps are used for Data pre-processing.

4.1.1 Normalization

The attribute data is scaled to fit into a specific range. There are many types of normalization available, but Min-Max Normalization will be useful. Min-Max Normalization transforms a value A to B which fits in the range [C, D]. It is given by the formula below

$$B = \frac{(A - minimum\ value\ of\ A)}{(maximum\ value\ of\ A - minimum\ value\ of\ A}) * (D - C) + C \ldots 1$$

Normalization technique and data pre-processing can really help get useful and right information about your data before applying some machine learning or data mining algorithm.

Normalizing the data attempts to give all attributes an equal weight. Normalization is particularly useful for classification algorithms.

4.1.2 Discretization

Discretization [11] is the process of converting the continuous domain of a feature into a nominal domain with a finite number of values. Front-end discretization might be necessary for some classifiers if their algorithms cannot handle continuous features by design. Additionally, earlier studies showed that discretization improves the accuracy of classifiers, including naïve Bayes classifiers, especially in larger datasets.

Numerous studies have examined discretization methods in the last two decades to determine how continuous values should be grouped, how cut points should be positioned on the continuous scale, and how many intervals should be used to generate datasets.

In this study, EWD i.e. Equal Width Discretization technique will be used. This method is selected because of their performance on large datasets, particularly the KDD'99 dataset. It is the simplest method to discretize a continuous-valued attribute by creating a specified number of bins. The bins can be created by equal-width.

In this method, k is used to determine the number of bins. Each bin is associated with a distinct discrete value. In equal-width, the continuous range of a feature is evenly divided into intervals that have an equal-width and each interval represents a bin. EWD divides the number line between Vmin and Vmax into k intervals of equal width; k is a user predefined parameter and usually is set as 10.

It divides the range into N intervals of equal size: uniform grid .if A and B are the lowest and highest values of the attribute, the width of intervals will be:

$$W = \frac{(B - A)}{N} \ldots 2$$

4.1.3 Feature Selection

Feature selection [12], [13] is a process of selecting a subset of relevant features by applying certain evaluation criteria. This approach removes redundant or irrelevant features from the dataset to prevent decreases in classification accuracy and unnecessary increases in computational costs. We are here using here Feature Subset selection method i.e. Correlation Based Feature Selection method (CFS Method.)

The CFS measure evaluates the subset of features based on the two concepts: feature-feature correlation and feature classification correlation [14]. The feature-feature correlation indicates the correlation between two features, while feature classification correlation says how much a feature is correlated to a specific class.

1) Correlation-based Feature Selection, CFS: Correlation based Feature Selection (CFS) is a simple filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function. The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and uncorrelated with each other. Irrelevant features should be ignored because they will correlation with the class. Redundant features should be screened out as they will be highly correlated with one or more of the remaining features. The acceptance of a feature will depend on the extent to which it predicts classes in areas of the instance space not already predicted by other features [15].

CFS's feature subset evaluation function is:

$$Merit_S = \frac{K\overline{r_{cf}}}{\sqrt{K + K(K-1)\overline{r_{ff}}}} \dots 3$$

Where MS is the heuristic 'merit' of a feature subset S containing k features, rcf is the mean feature-class correlation and rff is the average feature–feature inter correlation. The numerator of this equation can be thought of as providing an indication of how predictive of the class a set of features is; and the denominator of how much redundancy there is among the features.

4.2 Naïve Bayes

A Naïve Bayes (NB) classifier is a simple probabilistic classifier based on Bayes theorem where every feature is assumed to be class-conditionally independent. In naïve Bayes learning, each instance is described by a set of features and takes a class value from a predefined set of values. Classification of instances gets difficult when the dataset contains a large number of features and classes because it takes enormous numbers of observations to estimate the probabilities. When a feature is assumed to be class-conditionally independent, it really means that the effect of a variable value on a given class is independent of the values of other variables.

4.2.1 Bayesian Networks Classification

The Bayesian network is one of the most common classifiers for statistical data mining methods. The Bayesian network is based on a directed acyclic graph (DAG), where nodes represent attributes, and arcs represent attribute dependencies. In this method, the conditional probabilities for each node, which are based on its parents' attributes, quantify the attribute dependencies. A features, which consist of attributes {A1, A2, A2, …,Ai, …,An}, are represented as nodes in a Bayesian network, and (a1, a2, …,ai, …,an) are the attribute values of an instance Ei. The class variable C is represented as the top node in a Bayesian network, and c represents the value that C takes for instance E. The Bayesian network classifier can be defined as

$$c(E) = arg \max_{c \in C} P(c) \prod_{i=1}^{n} P(a_i|c) . \quad \dots 4$$

The naïve Bayes classification model is one of the most popular models because of its simplicity and computation efficiency, both of which are inherited from its conditional independence assumption property, as well as its good performance on datasets for which this property is fairly accurate. However, the model does not perform well if this assumption property is not satisfied, as observed in datasets that have complex attribute dependencies, such as the KDD'99 intrusion detection dataset

A Bayesian Network (BN) [16] is a graphical model for probability relationships among a set of variables features. The Bayesian network structure S is a directed acyclic graph (DAG) and the nodes in S are in one-to-one correspondence with the features X. The arcs represent casual influences among the features while the lack of possible arcs in S encodes conditional independencies. Moreover, a feature (node) is conditionally independent from its non-descendants given its parents (X1 is conditionally independent from X2 given X3 if P (X1|X2, X3) =P (X1|X3) for all possible values of X1, X2, X3).Typically, the task of learning a Bayesian network [17] can be divided into two subtasks: initially, the learning of the DAG structure of the network, and then the determination of its parameters. Probabilistic parameters are encoded into a set of tables, one for each variable, in the form of local conditional distributions of a variable given its parents. Given the independences encoded into the network, the joint distribution can be reconstructed by simply multiplying these tables. Within the general framework of inducing Bayesian networks, there are two scenarios: known structure and unknown structure. In the firstscenario, the structure of the network is given (e.g. by an expert) and assumed to be correct. Once the network structure is fixed, learning the parameters in the Conditional Probability Tables (CPT) is usually solved by

estimating a locally exponential number of parameters from the data provided. Each node in the network has an associated CPT that describes the conditional probability distribution of that node given the different values of its parents. In spite of the remarkable power of Bayesian Networks, they have an inherent limitation. This is the computational difficulty of exploring a previously unknown network.

Given a problem described by n features, the number of possible structure hypotheses is more than exponential in n. If the structure is unknown, one approach is to introduce a scoring function (or a score) that evaluates the "fitness" of networks with respect to the training data, and then to search for the best network according to this score. Several researchers have shown experimentally that the selection of a single good hypothesis using greedy search often yields accurate predictions. The most interesting feature of BNs, compared to decision trees or neural networks, is most certainly the possibility of taking into account prior information about a given problem, in terms of structural relationships among its features. This prior expertise, or domain knowledge, about the structure of a Bayesian network can take the following forms:

1. Declaring that node is a root node, i.e., it has no parents.
2. Declaring that node is a leaf node, i.e., it has no children.
3. Declaring that node is a direct cause or direct effect of another node.
4. Declaring that node is not directly connected to another node.
5. Declaring that two nodes are independent, given a condition-set.
6. Providing partial nodes ordering, that is, declare that a node appears earlier thananother node in the ordering

V. ANALYSIS

ALGORITHMS:

i. AD TREE:

An alternating decision tree (AD Tree) is a machine learning method for classification. It generalizes decision trees and has connections to boosting. Boosting is a machine learning meta-algorithm for reducing bias in supervised learning.

ii. NB TREE ALGORITHM:

1) For each attribute Xi , evaluate the utility, u(Xi), of a split onattribute Xi . For continuous attributes, a threshold is alsofound at this stage.

2) Let j = argmaxi (ui), i.e., the attribute with the highest utility.

3) If uj is not significantly better than the utility of the currentnode, create a Naive-Bayes classifier for the current node andreturn.

4) Partition the set of instances T according to the test on Xj . IfXj is continuous, a threshold split is used; if Xj is discrete, a multi-way split is made for all possible values.

5)For each child, call the algorithm recursively on the portion ofT that matches the test leading to the child.

iii. NAÏVE BAYES:

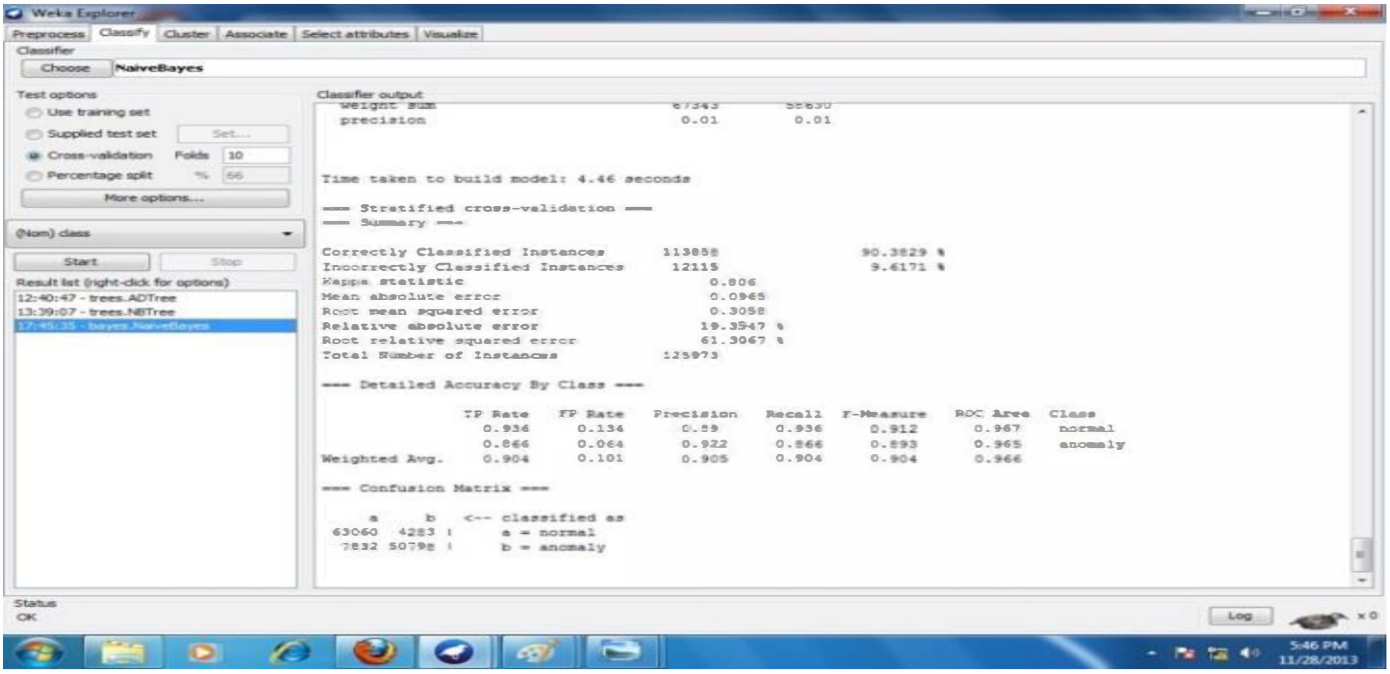

FIGURE 5.1: WEKA RESULTES FOR NAÏVE BAYES ALGORITHM

Figure 5.1 shows the output for Naïve Bayes algorithm by using Weka data mining tool. ForExperiments, we used Weka data mining tool for analyzing the results. The classification accuracy, TPRate, ROC and execution time is calculated for various classification algorithms such as NB TREE, Naïve Bayes and AD Tree as shown in Table 5.1.

TABLE 5.1:COMPARISON ANALYSIS FOR VARIOUS CLASSIFICATION ALGORITHMS

EVALUATION MEASURES	NB TREE	NAÏVE BAYES	AD TREE
CCI	0.998778	0.903829	0.984902
ICI	0.01222	0.096177	0.15098
PRECISION	0.999	0.89	0.983
TP RATE	0.999	0.936	0.989
FP RATE	0.002	0.134	0.019
ROC	1.000	0.967	0.998
EXECUTION TIME	1086.12 Sec	4.46 Sec	195.27 Sec

It can be observed from Figure 5.2 that the time taken by NB Tree to build the model is more compared to other classifiers and the time taken by naïve Bayes is less.

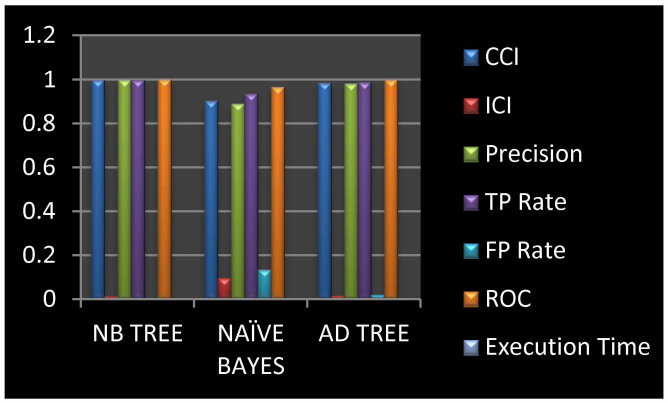

FIGURE5.2: COMPARISON ANALYSIS

Comparison on Evaluation Parameters

ACCURACY: The accuracy can be defined as the ratio of the number of correctly classified instances to the total number of examined instances.

FP RATE: FP Rate or False Positive Ratio/Rate Is the ratio of negative instances that were incorrectly classified as positive to the total number of negative instances.

TP RATE: TP Rate or True Positive Rate is the ratio of positive instances that were correctly identified to the total number of actual positive instances.

PRECISION: Precision Refers to the ratio of predicted positive instances that were correct to the total number of false positive and true positive.

RECALL: Recall Refers To The Part Of Relevant Information Which Is Actually Retrieved.

F MEASURE: On the basis of Precision and Recall the F MEASURE value is evaluated.

ROC AREA: Receiver operating curve is a plot with the FPR on the X-axis and the TPR on the Y-axis. This graph is very useful for measuring uniform classifier performance.

CONCLUSION

The proposed system explains the need to apply data mining methods to network events to classify network attacks. The paper defines a way to implement the use of naïve Bayes approach on the existing NSL-KDD 99 Dataset with the help of existing Discretization, Normalization and Feature selection methods.

After applying Normalization, Discretization and Feature selection method the proposed model will improve the results in terms of detection accuracy, error rate and misclassification cost.With respect to the TP Rate of all the algorithms presented in the paper it can be clearly visualized that the deviation in the execution time is less. It is also observed that time taken by NBTree to build the model is more compared to other classifier.Hence Naïve Bayes algorithm is efficient because of its simplicity, elegance, robustness and effectiveness as compared to other algorithms simulated in the paper.

REFERENCES

[1] Nguyen, HuyAnh, and Deokjai Choi. "Application of data mining to network intrusion detection: classifier selection model." Challenges for Next Generation Network Operations and Service Management. Springer Berlin Heidelberg, 2008.399-408.

[2] Dokas, P., Ertoz, L., Kumar, V., Lazarevic, A., Srivastava, J., & Tan, P. (2002). Data mining for network intrusion detection. In Paper presented at the proceedings of the nsf workshop on next generation data mining, Baltimore.

[3] Pages 13243-13252 Ambwani, T. (2003, 20-24 July 2003). Multi class support vector machine implementation to intrusion detection. Paper presented at the International Joint Conference on Neural Networks, 2003.

[4] Amudha, P., and H. Abdul Rauf. "Performance Analysis of Data Mining Approaches in Intrusion Detection." Process Automation, Control and Computing (PACC), 2011 International Conference on. IEEE, 2011.

[5] Kumar, Manish, M. Hanumanthappa, and T. V. Kumar. "Intrusion Detection System using decision tree algorithm." Communication Technology (ICCT), 2012 IEEE 14th International Conference on.IEEE, 2012.

[6] Kayacik, H. Günes, A. NurZincir-Heywood, and Malcolm I. Heywood. "Selecting features for intrusion detection: a feature relevance analysis on KDD 99 intrusion detection datasets."

Proceedings of the third annual conference on privacy, security and trust. 2005.

[7] Tavallaee, Mahbod, EbrahimBagheri, Wei Lu, and Ali-A. Ghorbani. "A detailed analysis of the KDD CUP 99 data set." In Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defense Applications 2009. 2009.

[8] Lee, Wenke, Salvatore J. Stolfo, and Kui W. Mok. "A data mining framework for building intrusion detection models. Security and Privacy, 1999.Proceedings of the 1999 IEEE Symposium on.IEEE, 1999.

[9] Olusola, Adetunmbi A., Adeola S. Oladele, and Daramola O. Abosede. "Analysis of KDD'99 Intrusion Detection Dataset for Selection of Relevance Features."Proceedings of the World Congress on Engineering and Computer Science.Vol. 1. 2010.

[10] MrutyunjayaPanda,ManasRanjanPatra,AComparative Study Of Data Mining Algorithms For Network Intrusion Detection, First International Conference on Emerging Trends in Engineering and Technology,2008 IEEE.

[11] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[12] Bolon-Canedo, V., Sanchez-Maroo, N., Alonso-Betanzos, A. (2009, 14-19 June). A combination of discretization and filter methods for improving classification performance in KDD Cup 99 dataset. Paper presented at the International Joint Conference on Neural Networks, IJCNN 2009,(pp. 14–19).

[13] Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2011). Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset. Expert Systems with Applications, 38(5), 5947–5957.

[14] Depren, O., Topallar, M., Anarim, E., &Ciliz, M. K. (2005). An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks. Expert Systems with Applications, 29(4), 713–722.

[15] Kabiri, P., Ghorbani, (2005). Research on intrusion detection and response: A survey. International Journal of Network Security, 1(2), 84–102.

[16] Beniwal, Sunita, and JitenderArora. "Classification and Feature Selection Techniques in Data Mining." International Journal of Engineering Research and Technology (IJERT) 1.6 (2012).

[17] Levent Koc , Thomas A. Mazzuchi, ShahramSarkani, A network intrusion detection system based on a Hidden Naïve Bayes Multiclass classifier, Expert Systems with Applications, 2012.