ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 4, April 2015*

# An Architecture for Network Intrusion Detection System based on DAG Classification

**Sunil Choudhary[1], Pankaj Dalal[2]**

M.Tech (SE) Scholar, Computer Science & Engineering, SITE, Nathdwara, India[1]

Associate Professor, Computer Science & Engineering, SITE, Nathdwara, India[2]

**Abstract**: Intrusion detection is an effective approach of dealing with problems in the area of network security. Rapid development in technology has raised the need for an effective intrusion detection system as the traditional intrusion detection method cannot compete against newly advanced intrusions. In this paper we proposed a feature based intrusion data classification technique. The reduced feature improved the classification of intrusion data. The reduction process of feature attribute performs by DAG function along with feature correlation factor. The proposed method work as feature reducers and classification technique, from the reduction of feature attribute also decrease the execution time of classification. For evaluation purposes, this model is applied to KDD '99 dataset.

**Keywords**: Network Intrusion Detection System, Directed acyclic graph, Classification, KDD'99 Data set, Support Vector Machine (SVM), Ensemble Technique, Neural Network.

## I. INTRODUCTION

Computer security is defined as the protection of computing systems against threats to confidentiality, integrity, and availability. Confidentiality means that information is disclosed only according to policy, integrity means that information is not destroyed or corrupted and that the system performs correctly, availability means that system services are available when they are needed. Security threats come from different sources such as natural forces, accidents, failure of services and people known as intruders. 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.

An intrusion detection system gathers and analyzes information from various areas within a computer or a network to recognize possible security breaches. Although, intrusion detection is the act of detecting actions that challenge to negotiation the confidentiality, integrity or availability of a system/network. Typically, intrusion detection systems have been categorized as a signature detection system, an anomaly detection system or a hybrid/compound detection system. A signature detection system identifies patterns of traffic or application data supposed to be malicious even though anomaly detection systems compare activities beside a "normal" baseline. Alternatively, a hybrid intrusion detection system combines the techniques of the two approaches.

A cyber attack (or intrusion) can be defined as a series of malicious computer activities that threaten and compromise the security and integrity of a computer/network system. The cyber attacks disrupt the normal operation of a computer system, and may illegally access or destroy the information in the computer systems. Most of the time, a cyber attack is launched through the data stream on the computer networks. The classification of cyber-attacks is helpful for learning the behavior of deferent attacks, which may be used in the design of cyber-attack detection systems. In general, cyber-attacks can be broadly classified into the following four categories:

1. Denial of Service (DoS) Attacks
2. Remote to Local (R2L) Attacks
3. User to Root (U2R) Attacks
4. Probing Attacks

Generally the cyber attack detection systems or intrusion detection systems (IDSs) are classified into two main categories:

1. Signature-based intrusion detection systems
2. Anomaly-based intrusion detection systems

### 1.1 SECURITY GOALS

There are five major security goals for network security. They are confidentiality, Availability, Authentication, Integrity and Non-repudiation.

#### A. Confidentiality

The information or data of any organization should be maintained in a safe manner and it should not be easily accessed by unauthorized users.

#### B. Availability

The information or data which plays a major role in an organization or in government offices should be stored secretly whereas it should be transparent to the authorized users and it should not be easily accessed by unauthorized users. It is necessary to fix up some limitations for the legitimate users.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 4, April 2015*

C. Authentication
The identity of the authorized users should be verified in order to access the information or data before the data is being accessed. There are three ways available to verify the identity of the legitimate user. They are password, tokens and biometrics.

D. Integrity
The information or data should not be altered during transmission.

E. Non-Repudiation
The sending and receiving parties of the information or data should ensure that both know about the delay in sending and receiving of the data or information. Apart from the primary goals of security there are certain other secondary goals that are required for maintaining security. They are access and availability.

1.2 CLASSIFICATION TECHNIQUES
Classification plays an important role in data mining and machine learning paradigm. The evaluation of ensemble classifier is great advantage over binary and conventional classifier. The process of prototype classification is combined two or more method with same nature. The prototype classification of data brings come in form of cluster oriented ensemble classifier. The need and requirement of online transaction of data is stream classification, due to stream classification save time of computation and storage area of network. For the purpose of stream data classification various machine learning algorithm are used, such as clustering, classification, and neural network. In the classification process of a growing data stream, also the temporary or long-standing activities of the stream may be more significant, or it often cannot be known a priori as to which one is more important. It is preferable that the window or horizon of the training data to use so as to obtain the best classification accuracy.

(i) Ensemble Classifier
Ensemble classification technique plays a vital role in data mining classification of data. The performance of individual classifier is not better in concern of accuracy and majority voting. The ensemble method started in last decade of machine learning research repository. When instances with known label are given the learning is called supervised learning and if instances are unlabelled the learning is called unsupervised learning. But unsupervised learning provides useful classes of items which is called clusters. Clusters are groups of similar types of objects. These groups are formed with classification methods [3]. For experimental purpose of COB three different ensemble methods bagging, random forests, and a randomized ensemble, two different numbers of individual classifiers and three different machine learning algorithms decision trees, k-nearest neighbours, and support vector machines are used.

1.3 TYPES OF MACHINE LEARNING
There are two major types of learning
A. Supervised Learning

B. Unsupervised Learning

In supervised learning the induced function is usually evaluated on a separate set of inputs and function values for them called the testing set. In supervised learning a function must be chosen to fit the training set from among a set of hypotheses.

A hypothesized function is said to generalize when it guesses well on the testing set. Curve fitting is a simple example of supervised learning. Inductive machine learning is the process of learning a set of rules from instances (examples in a training set), or more generally speaking, creating a classifier that can be used to generalize from new instances.

The process of applying supervised ML to a real-world problem is described in Figure 1.

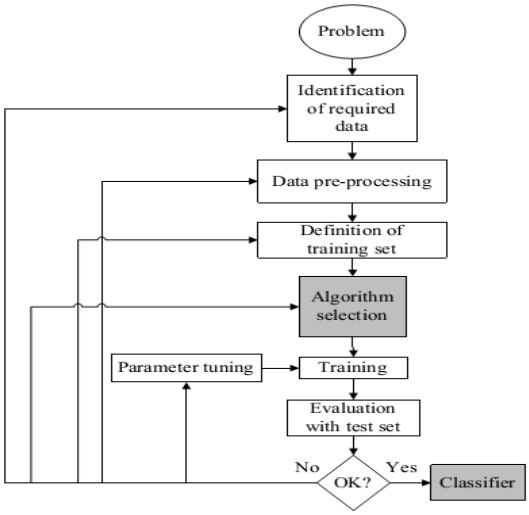


Figure 1: the process of supervised machine learning.

Supervised learning is one of the tasks most frequently carried out by so-called Intelligent Systems. Thus, a large number of techniques have been developed based on Artificial Intelligence (Logical/Symbolic techniques), Perception-based techniques and Statistics (Bayesian Networks, Instance-based techniques) [12].

In unsupervised learning training set of vectors are without function and they are partitioned into subset in some appropriate way. Unsupervised learning methods have application in taxonomic problems in which it is desired to invent ways to classify data into meaningful categories. Most of the unsupervised learning methods use a measure of similarity between patterns in order to group them into cluster [19]. Another type of unsupervised learning involves finding hierarchies of partitioning or clusters of clusters. A hierarchical partition is one in which is divided into mutually exclusive and exhaustive subsets, 1, ….., R and each subset are further divided into its mutually exclusive and exhaustive subsets.


DOI 10.17148/IJARCCE.2015.4447

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 4, April 2015*

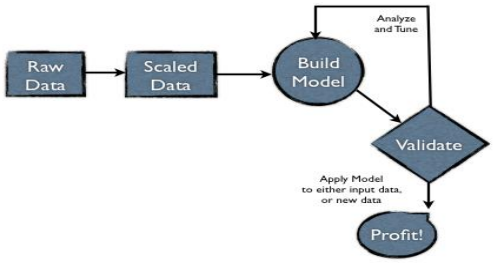


Figure 2: Unsupervised Learning

Classification problem is a problem where it is determined that whether an object is a member of a set or not. This problem can be divided into following categories [20]:

A. Binary classification
B. Multiclass classification
C. Statistical classification

Binary classification is the task of classifying the members of a given set of objects into two groups on the basis of whether they have some property or not. For a training set T with n training instances T = {(x1, y1), (x2, y2), …., (xn,yn)} , where each instance is a member of a problem domain xi ∈ Rm and a class label, yi ∈{c1,c2, …ck},where cj ≠ch for all h ≠j. The multi-class classification is a mapping function between instance X and class label Y where the number of K classes is greater than two, i.e. f: X →Y, K > 2. Generally, the multi-class classification problem is more difficult to handle than the binary classification problem [22]. This is because the number of classes could increase the complexity of the inductive learning algorithm.

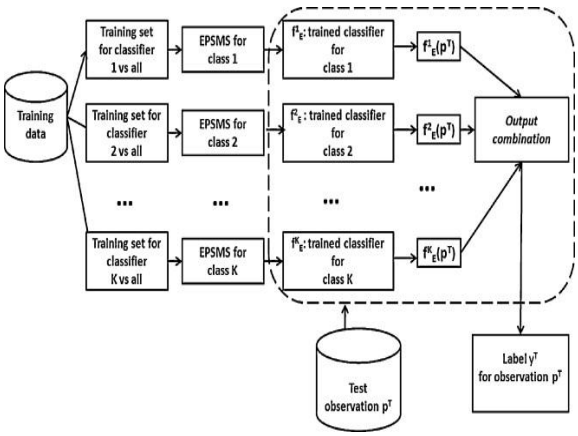


Figure 3: Multiclass Classification Approach

The data imbalance is another problem existed in the multiclass classification for machine learning. In the process of machine learning (ML) if the ration of minority class and majority class is highly different then machine learns more by the majority class and learns less from minority class. So to resolve the issue of minority class in imbalance data set special characteristic technique is required. Two more common approaches to solve this problem are data level approach and algorithm level approach. Data level approach rebalances the data before a classifier is trained and algorithm level approach strengthens the classifier. Alternatively, the learning algorithm can be modified to account for class imbalance. Cost-based learning is another such technique where instances in the minority class are assigned higher misclassification costs than those in the majority class.

## II. CLASSIFICATION AND ENSEMBLE TECHNIQUE

Classification is supervised learning process, by the nature of classification are divided into four section (1) Binary classification, (2) Rule-based classification, (3) Multi-class classification and finally (4) Neural network classification.

### 2.1 SAMPLING TECHNIQUES

Data level method for balancing the classes consists of resembling the original data set, either by oversampling the minority class or by under-sampling the majority class, until the classes are approximately equally represented. Both strategies can be applied in any learning system, since they act as a pre-processing phase, allowing the learning system to receive the training instances as if they belonged to a well-balanced data set. Thus, any bias of the system towards the majority class due to the different proportion of examples per class would be expected to be suppressed.

#### A. Oversampling

The simplest method to increase the size of the minority class corresponds to random over-sampling, that is, a non-heuristic method that balances the class distribution through the random replication of positive examples. Nevertheless, since this method replicates existing examples in the minority class, over fitting is more likely to occur.

#### B. Under sampling

Under-sampling is an efficient method for classing imbalance learning. This method uses a subset of the majority class to train the classifier. Since many majority class examples are ignored, the training set becomes more balanced and the training process becomes faster. Instances of the majority class are randomly discarded from the dataset. However, the main drawback of under-sampling is that potentially useful information contained in these ignored examples is neglected.

Sampling methods consider the class skew and properties of the dataset as a whole. However, machine learning and data mining often face nontrivial datasets, which often exhibit characteristics and properties at a local, rather than global level. It is noted that a classifier improved through global sampling levels may be insensitive to the peculiarities of different components or modalities in the data, resulting in a suboptimal performance. David A. [11] has suggested that for improving classifier performance sampling can be treated locally, instead of applying uniform levels of sampling globally. They proposed a framework which first identifies meaningful regions of data and then proceeds to find optimal sampling levels within each.


DOI 10.17148/IJARCCE.2015.4447

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 4, April 2015*

Oversampling is that it increases the number of training examples, thus increasing the learning time. Given the disadvantages with sampling, still sampling is a popular way to deal with imbalanced data rather than a cost-sensitive learning algorithm. There are several reasons for this. The most obvious reason is there are not cost sensitive implementations of all learning algorithms and therefore a wrapper-based approach using sampling is the only option.

2.2 BAGGING

Bagging (bootstrap aggregating) is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when expecting a class. Making bootstrap replicates of the learning sets forms the multiple versions [29]. Bagging can give substantial gains in accuracy when used on classification methods. Moreover, bagging can make unstable models stable and thus improve accuracy. A novel feature bagging approach for detecting outliers in very large data, and high dimensional and noisy databases.

2.3 BOOSTING

Boosting is a general method for improving (or boosting) the accuracy of any given learning algorithm. More precisely, boosting refers to a general and effective method of producing a very accurate prediction rule by combining rough and moderately inaccurate rules. By sequentially fitting models, later models see more of the samples misclassified by earlier ones. Combination uses weighted average where later models get more weight. Boosting works well on weak models and it reduces both bias and variance [24].

2.4 RANDOM FOREST

The random forest is an ensemble of unpruned classification or regression trees. Random forest generates many classification trees. Each tree is constructed by a different bootstrap sample from the original data using a tree classification algorithm. After the forest is formed, a new object that needs to be classified is put down each of the tree in the forest for classification. Each tree gives a vote that indicates the tree's decision about the class of the object. The forest chooses the class with the most votes for the object. The main features of the random forests algorithm are listed as follows:

- It is unsurpassable in accuracy among the current data mining algorithms.
- It runs efficiently on large data sets with many features.
- It can give the estimates of what features are important.
- It has no nominal data problem and does not over-fit.
- It can handle unbalanced data sets.

In random forests, there is no need for cross validation or a test set to get an unbiased estimate of the test error. Since each tree is constructed using the bootstrap sample, approximately one-third of the cases are left out of the bootstrap samples and not used in training. These cases are called out of bag (oob) cases. These oob cases are used to get a run-time unbiased estimate of the classification error as trees are added to the forest. Because random forests algorithm calculates the oob error during the training phase, there is no need to split the training data. It should choose the oob error estimate, since it is more effective by learning from the whole training dataset.

A. Bagging Trees

The basic idea underlying BT is the recognition that part of the output error in a single regression tree is due to the specific choice of the training data set. Therefore, if several similar data sets are created by resembling with replacement (that is, bootstrapping) and regression trees are grown without pruning and averaged, the variance component of the output error is reduced. When a bootstrap resample is drawn, about 37% of the data is excluded from the sample, but other data are replicated to bring the sample to full size. A related technique called ''boosting'' has recently gained popularity. In boosting, bias is reduced by repeatedly readjusting the weights of the training samples, by focusing on ''difficult'' examples from previous samples. Boosting is competitive with bagging and is used primarily in classifying data with large training sample sizes. Because it's primary goal was regression and not classification, it does not include boosting in their comparisons. Random Forests Random Forests is a new entry to the field of data mining and is designed to produce accurate predictions that do not over fit the data.

RF is similar to BT in that bootstrap samples are drawn to construct multiple trees; the difference is that the each tree is grown with a randomized subset of predictors, hence the name ''random'' forests. A large number of trees are grown, hence a ''forest'' of trees. The number of predictors used to find the best split a teach node is a randomly chosen subset of the total number of predictors. Support Vector Machine: Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. They belong to family of generalized linear classifiers. SVMs attempt to separate data into multiple classes (two in the basic case) though the use of a hyper-plane.

2.5 SUPPORT VECTOR MACHINE

Support Vector Machines (SVMs) are also a good candidate for intrusion detection systems which can provide real-time detection capability, deal with large dimensionality of data. SVMs plot the training vectors in high dimensional feature space through nonlinear mapping and labelling each vector by its class. The data is then classified by determining a set of support vectors, which are members of the set of training inputs that outline a hyper plane in the feature space. SVM is a technique for solving a variety of learning, Classification and prediction problems. The basic SVM deals with two-class problems in which the data are separated by a hyper plane defined


DOI 10.17148/IJARCCE.2015.4447

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 4, April 2015*

by a number of support vectors. Support vectors are a subset of training data used to define the boundary between the two classes. In situations where SVM cannot separate two classes, it solves this problem by mapping input data into high-dimensional feature spaces using a kernel function. In high-dimensional space it is possible to create a Hyper plane that allows linear separation. Compared with the ANN, the SVM have two advantages. Firstly, the global optimum can be derived. Secondly, the over fitting problem can be easily controlled by the choice of a suitable margin that separates the data. Categories of samples and measures belonging different classes decide the effects on the objective function. That assures eliminating or reducing the impact of the noise and outlier samples for the objective function.

## 2.6 NEURAL NETWORKS

An artificial neural network consists of a collection of processing elements that are highly interconnected and transform a set of inputs to a set of desired outputs. The result of the transformation is determined by the characteristics of the elements and the weights associated with the interconnections among them. By modifying the connections between the nodes the network is able to adapt to the desired outputs. Unlike expert systems, which can provide the user with a definitive answer if the characteristics which are reviewed exactly match those which have been coded in the rule base, a neural network conducts an analysis of the information and provides a probability estimate that the data matches the characteristics which it has been trained to recognize. A generic form of a neural network intrusion detector is presented in the below Figure 4. A classification rate can also be computed if the system is designed to perform attacks multi classification.

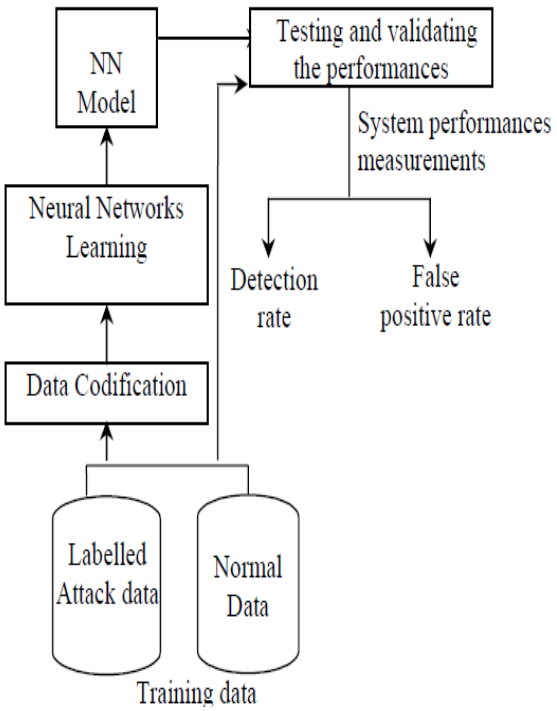


Figure 4 A generic form of a NN-base intrusion detection system

## 2.7 NETWORK DATA MINING

NDM may serve two different purposes. First, knowledge about the analyzed monitoring data is generated. This allows determining dominant characteristics and identifying outliers within the data records that can be considered as anomalous or suspicious. Secondly, NDM can be deployed to define rules or patterns that are typical for specific kinds of traffic, e.g. normal web traffic or traffic observed during a denial of service (DoS) attack. These rules and patterns can be used to analyze new sets of monitoring data and to check if these show similar properties and characteristics as the original data. Obvious applications profiting from such rules and patterns are network-based intrusion detection systems (NIDS) and traffic analyzers that characterize and classify traffic flows.

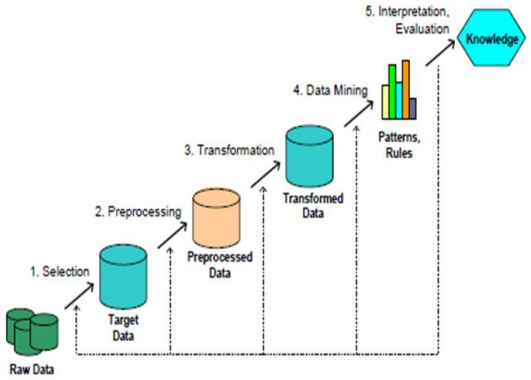


Figure 5. Knowledge Discoveries in Databases

The data mining approach to intrusion detection was first implemented in mining audit data for automated models for intrusion detection (MADAMID) The data mining process of building intrusion detection models is depicted in below figure 3.4. Raw data is first converted into ASCII network packet information, which in turn is converted into connection level information. These connection level records contain within connection features like service. Data mining algorithms are applied to this data to create models to detect intrusions. Data mining algorithms used in this approach include rule based classification algorithm (RIPPER), Meta classifier, frequent episode algorithm and association rules.

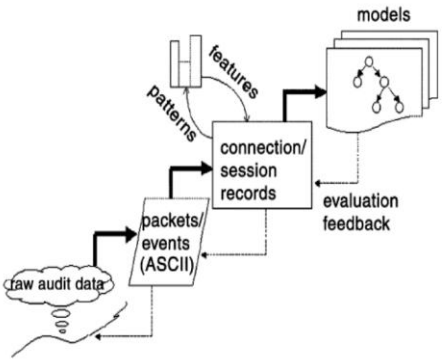


Figure 6. Data mining process of building Intrusion detection models.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 4, April 2015*

2.8 FALSE POSITIVES AND FALSE NEGATIVES

In order to evaluate the IDS's performance and detection accuracy there are four possible occurrences whose ratio is monitored. False positives are legal occurrences that are incorrectly marked as anomalous. True positives are occurrences that are correctly marked as anomalous. False negatives are anomalous occurrences that are missed by the detector and therefore are not marked as anomalous. True negatives are occurrences that are correctly marked as legal activity. In order to find out whether the anomaly or intrusion is a false positive or false negative, it has to be investigated by a network operator. [10]

2.9 KDD CUP 99

KDD CUP 1999 data was the data set used for the Third International Knowledge Discovery and Data Mining Tools Competition. The training data set contains 494,021 connection records, and the test data set contains 311,029 records. For this dissertation, it uses sample the data only from the training data set and use in both the training and testing stages. A connection is a sequence of TCP packets containing values of 41 features and labeled as either normal or an attack, with exactly one specific attack type. A complete listing of Features and details are in KDD CUP 1999 data. There are 22 attack types in the training dat. the attacks in the training data are grouped into broad classes; each neuron is then labeled as representing one of these classes. Specifically, the four broad classes of attack type defined by MIT Lincoln Labs are used, as stated below:

1) Denial-of-Service (DoS): These are attacks designed to make some service accessible through the network unavailable to legitimate users.
2) Probe: A Probe is a reconnaissance attack designed to uncover information about the network, which can be exploited by another attack.
3) Remote-to-Local (R2L): This is where an attacker with no privileges to access a private network attempts to gain access to that network from outside, e.g. over the internet.
4) User-to-Root (U2R): The attacker has a legitimate user account on the target network. However, the attack is designed to escalate his privileges so that he can perform unauthorized actions on the network.

## III. PROPOSED METHODOLOGY AND ARCHITECTURE

Intrusion data classification and detection process is very complex process in network security. In current network security scenario various types of Intrusion attack are available some are known attack and some are unknown attack. The attack of know Intrusion detection used some well know technique such as signature based technique and rule based technique. In case of unknown Intrusion attack detection of attack is various challenging task. All paragraphs must be indented.

3.1 FEATURES EXTRACTION

Intrusion classification can either have single variable approach or a multi-variable approach to detect Intrusion depending on the algorithm used. In the single variable approach a single variable of the system is analyzed. This can be, for example, port number, CPU usage of a local machine etc. In multi-variable approach a combination of several features and their inter-correlations are analyzed. [24] In addition based on the method the way in which features are chosen for the IDS can be divided into two groups; into feature selection and feature reduction.

A. Feature Selection

In the feature selection method the features are either picked manually from the data monitored or by using a specific feature selection tool. The most suitable features are selected by handpicking from the feature spectrum based on the prior knowledge about the environment that the IDS are monitoring. For example features that can distinguish certain type of traffic from the traffic flows are picked for the network traffic model training. The idea behind the feature selection tools is to reduce the amount of features into a feasible subset of features that do not correlate with each other. Feature selection process is illustrated in Figure 4.1 On the left there are the features (F0…FN) that are available from the data monitored, which is, for example, from network traffic. On the right side is the output (F0...FM) of the selection tool. The number of features in the output varies based on the selection tool used and the inter-correlation of features in the input. Following the basic principles of feature analysis the number of features in the output (M in Figure 4.1) is in most of the cases less than the number of features in the input (N in Figure 7). However, it is possible that the output is equal to the input.

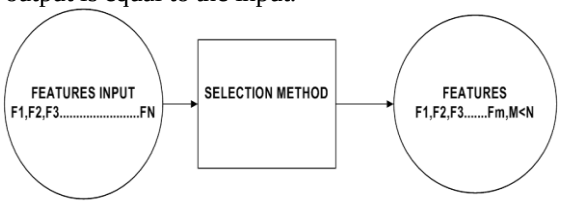


Figure 7. Feature selection process in feature variable

3.2 SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) is a novel machine learning method based on statistical learning theory developed by V.N.Vapnik, and it has been successfully applied to numerous classification and pattern recognition problems such as text categorization, image recognition and bioinformatics. It is still in the development stage now [14]. SVM can be used for pattern recognition, regression analysis and principal component analysis. The achievements of SVM in training have Platt's the sequential minimal optimization method, Osuna's the method of Chunking, Joachims' SVM light method and so on [16]. These methods are directed at the training process, and not related to classification process. In the process of SVM training, all the samples are used. So it has no effect on the speed of the classification. Lee and others propose a method of reduction SVM training time and adding the speed of training, reduced support vector machines [20]. Burges put forward a way of increasing the speed of Classification ,which does not use the support

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 4, April 2015*

vector in the category function but use a reduction of vector set, which is different from the standard vector set .That is neither training samples nor support vector but it is the transformation of the special vector. The method achieved certain results, but in the process of looking for the reduction of the vector collection, the cost of calculation paid is too large to widely use in practice. The concept of SVM is to transform the input vectors to a higher dimensional space Z by a nonlinear transform, and then an optical hyperplane which separates the data can be found. This hyperplane should have the best generalization capability. As shown in Figure 8, the black dots and the white dots are the training dataset which belong to two classes. The Plane H series are the hyperplanes to separate the two classes. The optical plane H is found by maximizing the margin value . Hyperplanes and are the planes on the border of each class and also parallel to the optical hyperplane H. The data located on and are called support vectors.

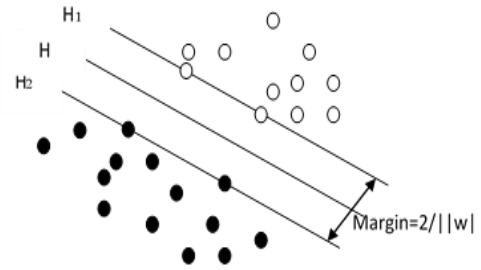


Figure 8. The SVM binary classifications

For training data set, to find the optical hyperplane H, a nonlinear transform, is applied to x, to make x become linearly dividable. A weight and offset satisfying the following criteria will be found:

$$\begin{cases} w^T z_i + b \geq 1, & y_i = 1 \\ w^T z_i + b \leq -1, & y_i = -1 \end{cases} \quad \text{…….. (i)}$$

The development of a SVM image classification model depends on the selection of kernel function K. There are several kernels that can be used in Support Vector Machines models. These include linear, polynomial, radial basis function (RBF) and sigmoid function:

$$K(x_i, x_j) = \begin{cases} x_i^T x_j & Linear \\ (\gamma x_i^T x_j + coefficient)^{degree} & Polynomial \\ \exp(-\gamma \mid x_i - x_j \mid^2 & RBF \\ \tanh(\gamma x_i^T x_j + coefficient) & Sigmoid \end{cases} \quad \text{…… (ii)}$$

The RBF is by far the most popular choice of kernel types used in Support Vector Machines. This is mainly because of their localized and finite responses across the entire range of the real x-axis.

Improper kernel function might generate poor performance. Currently there is no effective "learning" method to choose a proper kernel function for a specific problem. The selection is decided by the experiment result at this time. In our proposed system, two kernel functions are tested: Radial Basis Function-RBF and Polynomial Function.

$$K_{poly}(x_1, x_2) = (x_1 * x_2 + 1)^p \quad \text{……. (iii)}$$

$$K_{RBF}(x_1, x_2) = \exp(-p \parallel x_1 - x_2 \parallel^2) \quad \text{…….. (iv)}$$

Due to its better performance, RBF was chosen as the kernel function in the model.

### 3.3 DIRECTED ACYCLIC GRAPH (DAG)

A DAG is a graph based multi-classification technique in this technique pair-wise SVMs used, let the decision function for class *i*against class *j*, with the maximal margin, be:

$$Dij(x) = wijT\emptyset(x) + bij \quad \text{…….. (v)}$$

And x is classified into class

$$\arg \max_{i = 1,2,\ldots,n} Di(x) \quad \text{….. (vi)}$$

If xϵRiDi(x)=n-1 and $D_k(x)$<n-1 for k≠i. Thus x is classified into i. But if any of Di(x) is not n-1 may be satisfied for plural iS. In this case x is unclassified. In the shaded region in figure 9,Di(x)=0(i=1,2 and 3). Therefore, this region is unclassified, although the unclassified region is much smaller than that for the one-against-all support vector machine.

In pairwise SVMs, classification reduces the unclassifiable regions that occur for one-against-all support vector machines but it still exists. To resolve this problem, Vapnik [2] proposed to use continuous decision functions. Namely, we classify a datum into the class with maximum value of the decision functions. Inoue and Abe [21] proposed fuzzy support vector machines, in which Membership functions are defined using the decision functions.Another popular solution is DAG SVM that uses a decision tree in the testing stage. Training of a DAG is the same as conventional pairwise SVMs.

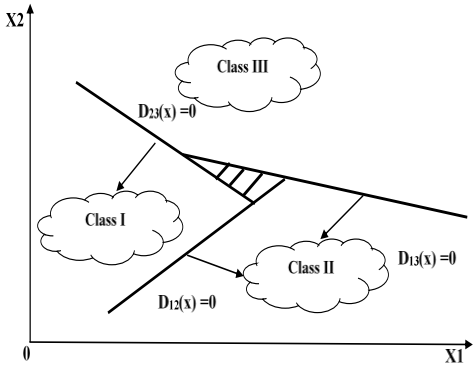


Figure 9: Unclassifiable regions by the pairwise formulation.

Classification by DAGs is faster than by conventional pairwise SVMs or pairwise fuzzy SVMs. Figure 10 shows the decision tree for the three classes shown in Figure 9. In the figure 10, i show that x does not belong to class i.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 4, April 2015*

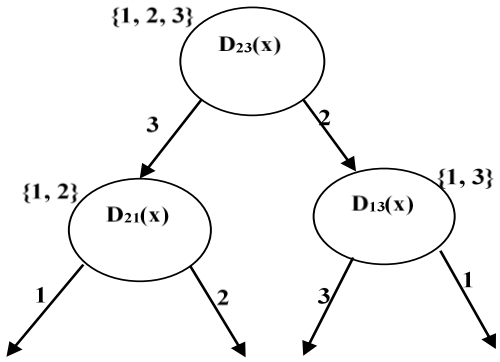


Figure 10: DAG classification.

As the top-level classification, it can choose any pair of classes. And except for the leaf node if Dij(x)>0, lets consider that x does not belong to class j, and if Dij(x)<0 not class i.

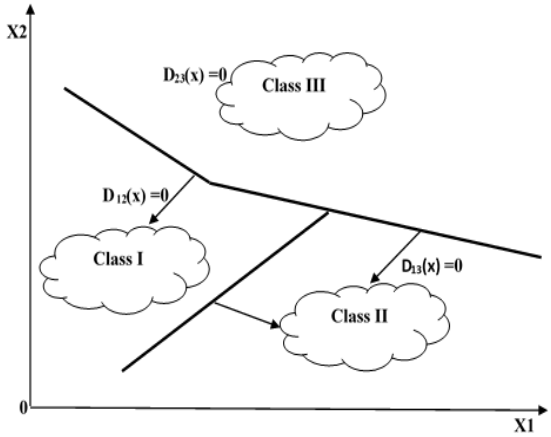


Figure 11: Generalization region by DAG classification.

Thus, if $D_{12}(x)> 0$, x does not belong to class II. Therefore, it belongs to either class I or class III, and the next classification pair is classes I and III. The generalization regions become as shown in Figure 11. Unclassifiable regions are resolved, but clearly the generalization regions depend on the tree formation.

3.4 PROPOSED METHOD

In this section discuss the graph based ensemble based intrusion detection technique. The graph based technique basically work with collection of features of attribute of network data. The network traffic data passes through the graph set, set collect the similar type of attribute and discard the dissimilar attribute.

The discarded attribute used as feature of collection attribute for ensemble process. The ensemble point fist select the all feature attribute for selection and pass through the classification.

The process of algorithm discuss in two phase first phase discuss the collection of attribute and second phase discuss the ensemble of attribute for classification.

**Phase-I**

**Step1:** Initially input Intrusion data passes through preprocessing function and extracted feature part of Intrusion data in form of traffic type.

**Step2:** the extracted traffic feature data converted into feature vector.

**Step 3:** In phase of feature mapping in feature space of DAG create a fixed class according to the group of data.

**Step 4:** steps of processing of DAG.

1) Initialize Gaussian hyper plane margin.
2) Choose a random vector from training data and present it to the DAG.
3) The weight of the plane support vector is estimated. The size of the vector decreases with each iteration.
4) Each vector in the SV's neighborhood has its weights adjusted to become more like the SV. Vector closest to the SV are altered more than the vector furthest away in the neighborhood.
5) Repeat from step 2 for enough iteration for convergence.
6) Calculating the SV is done according to the Euclidean distance among the node's weights ($W_1$, $W_2$…, $W_n$) and the input vector's values ($V_1$, $V_2$, …. , $V_n$).
7) The new weight for a node is the old weight, plus a fraction (L) of the difference between the old weight and the input vector… adjusted (theta) based on distance from the SVM

**Phase-II**

Input: N_list: collection of intrusion attributes
Output: N_type: number of classified class

1) G = (V, E) ←empty //define the feature data in graph mode
2) NP_list←K-means(N_list, $K_v$) //grouping of data
3) for h ∈ N P_list do
4) h.nn ← Nearest-neighbor (NP_list - {h})
5) h.sc←Compute-SC(h,h.nn) //Reduction of attribute
6) V←V ∪ {h} //commutate number of attribute
7) V←V ∪{h.nn}
8) if h.sc < $th_{sc}$ then //check class group
9) E←E ∪ {(h,h.nn)} //add this DAG
10) endif
11) end for
12) for each pair of components (g1,g2) ∈ G do
13) $\mu_1$←mean-dist (g1), $\mu_2$←mean-dist (g2)
14) If $\frac{\mu_1+\mu_2}{2*centroid\_dist(g1,g2)}$ > 1 then g1←Merge (g1, g2)
15) end for // Now allot the class labels
16) N_type ← empty
17) for $x$ ∈ N list do
18) h ← PseudopointOf ($x$)
19) N_type ← N_type ∪ {($x$) , h.ccomponent )}
20) end for

**Step 5:** After processing of support vector finally Intrusion data are classified.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 4, April 2015*

3.5 PROPOSED MODEL

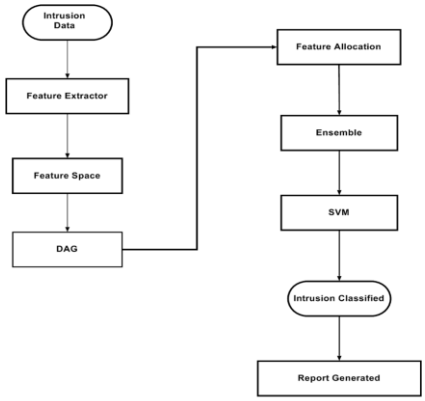


Figure 12 Proposed Model for Intrusion classification

## IV. IMPLEMENTATION & RESULT ANALYSIS

Total sample description of Data Set attacks of four types of attack. A variety of attacks incorporated in the dataset fall into following four major categories:

**Denial of Service Attacks:** A denial of service attack is an attack where the attacker constructs some computing or memory resource fully occupied or unavailable to manage legitimate requirements, or reject legitimate users right to use a machine.

**User to Root Attacks:** User to Root exploits are a category of exploits where the attacker initiate by accessing a normal user account on the system (possibly achieved by tracking down the passwords, a dictionary attack, or social engineering) and take advantage of some susceptibility to achieve root access to the system.

**Remote to User Attacks:** A Remote to User attack takes place when an attacker who has the capability to send packets to a machine over a network but does not have an account on that machine, makes use of some vulnerability to achieve local access as a user of that machine.

**Probes:** Probing is a category of attacks where an attacker examines a network to collect information or discover well-known vulnerabilities. These network investigations are reasonably valuable for an attacker who is staging an attack in future. An attacker who has a record, of which machines and services are accessible on a given network, can make use of this information to look for fragile points.

A. Different Types of Attacks In Kdd99 Dataset

| 4 Main Attack Classes | 22 Attacks Classes |
|---|---|
| Probing | ipsweep, nmap, portsweep, satan |
| Denial of Service (DOS) | back, land, eptune, pod, smurt, teardrop |
| User to Root (U2R) | buffer_overflow, perl, loadmodule, rootkit |
| Remote to User (R2L) | ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster |

TABLE 1: TYPES OF ATTACK IN KDD 99 DATASET

On experimenting with different dataset, the number of normal/abnormal packets is being monitor. We have examined five different dataset in our experiment, with each having corresponding number of rejected or normal packets. In our conducted test the packets could either fall under normal packet type or in the category of attack (DOS, R2L.U2R.PROB).

We have supervised on data set with each 7000 instances of data under .the result of predicted normal and abnormal data is form of confusion matrix.

TP: True Positive
TN: True Negative
FP: False Positive
FN: False Negative

B. PERFORMANCE PARAMETERS

$$\text{PRECISION} = \frac{TP}{TP+FP}$$

$$\text{RECALL} = \frac{TP}{TP+FN}$$

$$\text{ACCURACY} = \frac{TP+TN}{TP+TN+FN+FP}$$

$$\text{FPR} = \frac{FP}{FP+TN} \quad , \text{FNR} = \frac{FN}{FN+TP}$$

4.1 PERFORMANCE EVALUEATION

| Method Name | Value | Types Of Attack | TPR | TNR | FPR | FNR | Detection Rate | Precision Rate | Recall Rate |
|---|---|---|---|---|---|---|---|---|---|
| Ensemble | 0.1 | Normal | 4.273 | 0.703 | 1.568 | 0.691 | 89.79 | 81.93 | 80.93 |
| | | DOS | 4.373 | 0.296 | 0.568 | 0.308 | 88.79 | 80.83 | 78.46 |
| | | PROBE | 4.483 | 1.703 | 0.191 | 0.351 | 86.79 | 79.56 | 81.93 |
| | | U2R | 5.273 | 0.407 | 1.431 | 0.308 | 85.79 | 82.93 | 81.93 |
| | | R2L | 3.473 | 1.592 | 0.568 | 0.168 | 86.79 | 84.43 | 79.93 |

TABLE 2: SHOWS THAT THE PERFORMANCE EVALUATION OF TPR, TNR, FPR, FNR, DETECTION RATE, PRECISION RATE AND RECALL RATE FOR ENSEMBLE METHOD, AND THE INPUT VALUE IS 0.1.

| METHOD NAME | VALUE | TYPES OF ATTACK | TPR | TNR | FPR | FNR | DETECTION RATE | PRECISION RATE | RECALL RATE |
|---|---|---|---|---|---|---|---|---|---|
| ENSEMBLE | 0.5 | NORMAL | 6.032 | 2.462 | 3.327 | 2.450 | 91.55 | 83.69 | 82.69 |
| | | DOS | 6.123 | 1.564 | 2.354 | 1.450 | 90.55 | 82.69 | 80.69 |
| | | PROBE | 6.243 | 3.462 | 1.567 | 2.110 | 88.55 | 81.87 | 83.67 |
| | | U2R | 7.032 | 1.351 | 0.327 | 1.450 | 87.55 | 84.47 | 83.69 |
| | | R2L | 5.232 | 3.351 | 2.327 | 1.590 | 88.55 | 86.19 | 81.68 |

TABLE 3: SHOWS THAT THE PERFORMANCE EVALUATION OF TPR, TNR, FPR, FNR, DETECTION RATE, PRECISION RATE AND RECALL RATE FOR ENSEMBLE METHOD, AND THE INPUT VALUE IS 0.1.

| Method Name | Value | Types Of Attack | TPR | TNR | FPR | FNR | Detection Rate | Precision Rate | Recall Rate |
|---|---|---|---|---|---|---|---|---|---|
| IMPROVED ENSEMBLE | 0.1 | NORMAL | 3.654 | 0.843 | 1.748 | 0.751 | 95.80 | 85.02 | 83.97 |
| | | DOS | 3.513 | 1.843 | 1.738 | 0.741 | 93.83 | 81.97 | 80.97 |
| | | PROBE | 2.313 | 1.853 | 0.738 | 1.131 | 94.83 | 84.97 | 81.97 |
| | | U2R | 3.543 | 0.854 | 0.853 | 1.851 | 95.67 | 85.97 | 84.97 |
| | | R2L | 3.093 | 0.698 | 0.408 | 1.846 | 92.83 | 86.94 | 82.94 |

TABLE 4: SHOWS THAT THE PERFORMANCE EVALUATION OF TPR, TNR, FPR, FNR, DETECTION RATE, PRECISION RATE AND RECALL RATE FOR IMPROVED ENSEMBLE METHOD, AND THE INPUT VALUE IS 0.1.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 4, April 2015*

| Method Name | Value | Types Of Attack | TPR | TNR | FPR | FNR | Detection Rate | Precision Rate | Recall Rate |
|---|---|---|---|---|---|---|---|---|---|
| ENSEMBLE | 0.5 | NORMAL | 6.032 | 2.462 | 3.327 | 2.450 | 91.55 | 83.69 | 82.69 |
| | | DOS | 6.123 | 1.564 | 2.354 | 1.450 | 90.55 | 82.69 | 80.69 |
| | | PROBE | 6.243 | 3.462 | 1.567 | 2.110 | 88.55 | 81.87 | 83.67 |
| | | U2R | 7.032 | 1.351 | 0.327 | 1.450 | 87.55 | 84.47 | 83.69 |
| | | R2L | 5.232 | 3.351 | 2.327 | 1.590 | 88.55 | 86.19 | 81.68 |

TABLE 5: SHOWS THAT THE PERFORMANCE EVALUATION OF TPR, TNR, FPR, FNR, DETECTION RATE, PRECISION RATE AND RECALL RATE FOR ENSEMBLE METHOD, AND THE INPUT VALUE IS 0.5.

| Method Name | Value | Types Of Attack | TPR | TNR | FPR | FNR | Detection Rate | Precision Rate | Recall Rate |
|---|---|---|---|---|---|---|---|---|---|
| IMPROVED ENSEMBLE | 0.5 | NORMAL | 5.259 | 2.602 | 3.507 | 2.510 | 97.56 | 86.78 | 85.73 |
| | | DOS | 5.272 | 3.603 | 3.497 | 2.500 | 95.58 | 83.74 | 82.73 |
| | | PROBE | 4.072 | 3.612 | 2.497 | 2.890 | 96.57 | 86.81 | 83.7473 |
| | | U2R | 5.267 | 2.643 | 2.612 | 3.610 | 96.64 | 87.73 | 86. |
| | | R2L | 4.852 | 2.457 | 2.167 | 3.547 | 94.58 | 88.56 | 84.67 |

TABLE 6: SHOWS THAT THE PERFORMANCE EVALUATION OF TPR, TNR, FPR, FNR, DETECTION RATE, PRECISION RATE AND RECALL RATE FOR IMPROVED ENSEMBLE METHOD, AND THE INPUT VALUE IS 0.5.

## 4.2 RESULT ANALYSIS

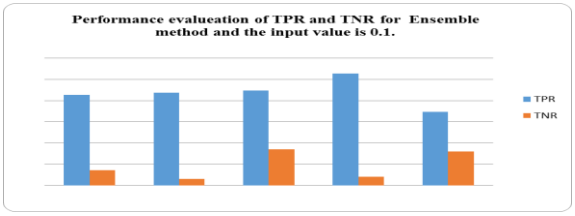


Figure 13. Shows that the performance evalueation of TPR and TNR for the ensemble method and the input value is 0.1.

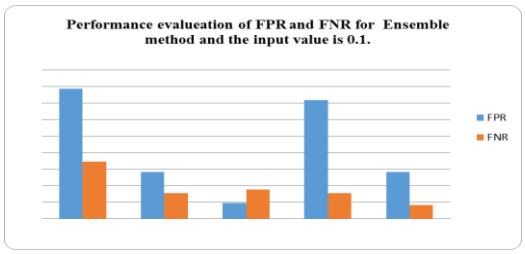


Figure14. Shows that the performance evalueation of FPR and FNR for the ensemble method and the input value is 0.1.

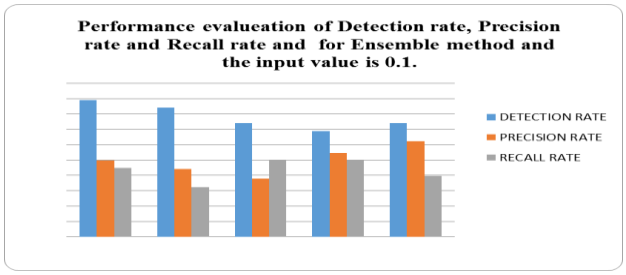


Figure 15. Shows that the performance evalueation of Detection rate, Precision rate and Recall rate for the ensemble method and the input value is 0.1.

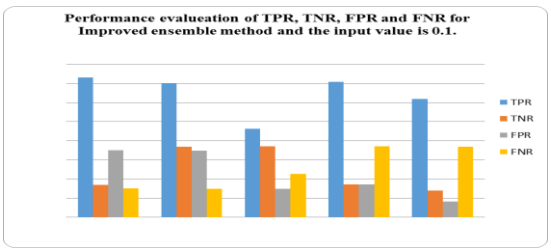


Figure 16. Shows that the performance evalueation of TPR, TNR, FPR and FNR for the Improved ensemble method and the input value is 0.1.

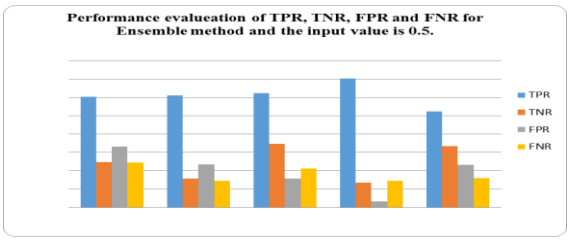


Figure 17. Shows that the performance evalueation of TPR, TNR, FPR and FNR for the ensemble method and the input value is 0.5.

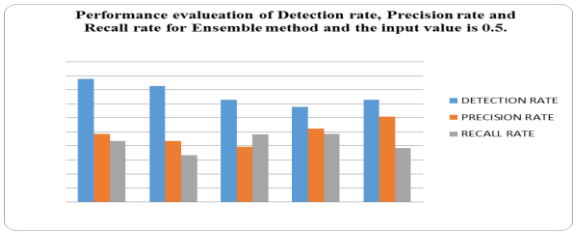


Figure 18. Shows that the performance evalueation of Detection rate, Precision rate and Recall rate for the ensemble method and the input value is 0.5.

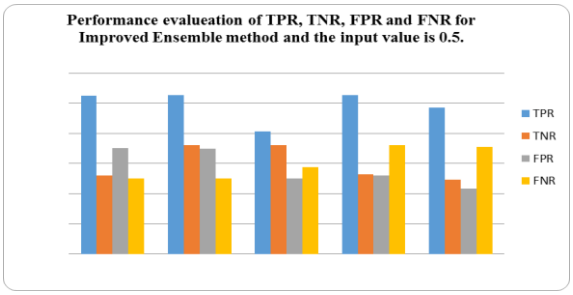


Figure 19. Shows that the performance evalueation of TPR, TNR, FPR and FNR for the Improved ensemble method and the input value is 0.5.

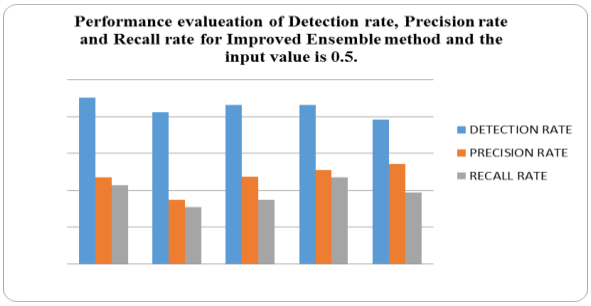


Figure 20. Shows that the performance evalueation of Detection rate, Precision rate and Recall rate for the Improved ensemble method and the input value is 0.5.


DOI 10.17148/IJARCCE.2015.4447

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 4, April 2015*

## V. CONCLUSION

The proposed method work as feature reducers and classification technique, from the reduction of feature attribute also decrease the execution time of classification. The decrease time increase the performance of intrusion detection system. Our experimental process gets some standard attribute set of intrusion file such as pot_type, service, sa_srv_rate, dst_host_count, dst_host_sa_srv_rate. These feature attribute are most important attribute in domain of traffic area. The classification rate in these attribute achieved 98 %. In this paper reduction computational time of feature selection process is main objective. Because of this consumed time of each algorithm with different reject threshold measured. As evaluation result shows, although FFR cannot defeat other methodologies in accuracy of classification and accuracy didn't changed very much, but in speed FFR outperformed all other feature selection method with great differences. We used ensemble classifier for developing efficient and effective IDS. For improving the detection rate of the minority classes in imbalanced training dataset we used standard sampling and we picked up all of the important features of the minority class using the minority classes attack mode.

## REFERENCES

[1]. Shafigh Parsazad, Ehsan Saboori, Amin Allahyar "Fast Feature Reduction in Intrusion Detection Datasets" MIPRO 2012, Pp 1023-1029.
[2]. Abebe Tesfahun, D. Lalitha Bhaskari "Intrusion Detection using Random Forests Classifier with SMOTE and Feature Reduction" International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, 2013. Pp 127-132.
[3]. Hachmi Fatma, Limam Mohamed "A two-stage technique to improve intrusion detection systems based on data mining algorithms" IEEE, 2013. Pp 1-6.
[4]. Shailendra Singh, Sanjay Silakari "An Ensemble Approach for Cyber Attack Detection System: A Generic Framework" 14th ACIS, IEEE 2013.
[5]. Li, "Using Genetic Algorithm for Network Intrusion Detection" Proc. the United States Department of Energy Cyber Security Group 2004 Training Conference, May 2004.
[6]. Jain , Upendra "An Efficient intrusion detection based on Decision Tree Classifier using feature Reduction", International Journal of scientific and research Publications , Vol. 2, Jan. 2012.
[7]. Dewan Md. Farid, Jerome Darmont, Nouria Harbi, Nguyen Huu Hoa, Mohammad Zahidur Rahman "Adaptive Network Intrusion Detection Learning: Attribute Selection and Classification" 2008. Pp 1-5.
[8]. Gary Stein, Bing Chen, Annie S. Wu, Kien A. Hua "Decision Tree Classifier For Network Intrusion Detection With GA-based Feature Selection" 2556. Pp 1-6.
[9]. Ritu Ranjani Singh a, Prof. Neetesh Gupta "To Reduce the False Alarm in Intrusion Detection System using self Organizing Map" in International journal of Computer Science and its Applications.
[10]. Z. Xue-qin, G. Chun-hua, L. Jia-jin "Intrusion detection system based on feature selection and support vector machine" Proc. First International Conference on Communications and Networking in China (ChinaCom '06), Oct. 2006.
[11]. Zhang , M. Zulkernine "Network Intrusion Detection using Random Forests" School of Computing Queen's University, Kingston Ontario, 2006.
[12]. John Zhong Lei and Ali Ghorbani "Network Intrusion Detection Using an Improved Competitive Learning Neural Network" in Proceedings of the Second Annual Conference on Communication Networks and Services Research IEEE.
[13]. P. Jongsuebsuk,N. Wattanapongsakorn and C. Charnsripinyo "Network Intrusion Detection with Fuzzy Genetic Algorithm for Unknown Attacks" in IEEE 2013.
[14]. Deepak Rathore and Anurag Jain "a novel method for intrusion detection based on ecc and radial bias feed forword network" in Int. J. of Engg. Sci. & Mgmt. (IJESM), Vol. 2, Issue 3: July-Sep.: 2012.
[15]. Wing w. Y. Ng, rocky k. C. Chang and daniel s. Yeung "dimensionality reduction for denial of service detection problems using rbfnn output sensitivity" in Proceedings of the Second International Conference on Machine Learning and Cybernetics, Wan, 2-5 November 2003.