ISSN:0975-9646

Advanced Network Intrusion Detection System Based on Effective Feature Selection

Sumathi M[1], Umarani R[2]

[1]*Department Of Computer Science, Mahendra Arts & Science College, Salem, Tamilnadu, India*

[2]*Department Of Computer Science,Sri Saradha College For Women, Salem-16, Tamilnadu, India*

Abstract-The growing rate of network attacks including hacker, cracker, and criminal enterprises have been increasing, which impact to the availability, confidentiality, and integrity of critical information data. Intrusion detection system have become a necessary addition to security infrastructure of most Organizations. Intrusion detection systems (IDS) take either network or host based approach for recognizing and deflecting attacks. Machine Learning includes a number of advanced statistical methods for handling regression and classification tasks.Methods include Support Vector Machines (SVM) for regression and classification, Naive Bayes for classification, and k-Nearest Neighbors (KNN) for regression and classification. Feature selection is a process that selects a subset of original features. Proposed a useful pre-processing step is to run your data through the following data cleaning routines.The proposed system is to enhance the Feature Selection process by using a new MLP based Feature Selection method. Supervised learning algorithms employ a collection of instances to estimate the class label of new input samples. The instances are generally represented by a number of attributes and a class value.In fact, some features can be redundant or irrelevant. By removing such redundant and irrelevant attributes, a classifier with higher predictive accuracy can often be obtained. The experimental results show better results and prediction accuracy.

Keywords
IDS,RST,SVM,PCA,KNN,MLP

1. INTRODUCTION

Data mining (DM) deals with the problem of discovering novel and interesting knowledge from large amount of data. Problem is often performed heuristically when the extraction of patterns is difficult using standard query mechanisms or classical statistical methods. A comparison with the results achieved by other techniques on a classical benchmark set is carried out. The core of this process is the application of machine learning based algorithms to databases. Two basic ways of performing data mining and they are supervised and unsupervised learning. In unsupervised learning, data patterns are found from some logical characterization of the regularities in a set of data. In this case, no pre-assumptions are made about the forms of relations among attributes. Data classification represents the most commonly applied supervised data mining technique.Data mining sometimes referred as Knowledge Discovery in Database (KDD), is a systematic approach to find the underlying patterns, trend and relationships buried in data. Researches regarding DM can be classified into two categories such as methodologies and technologies. The technology part of DM consists of techniques such as statistical methods, neural networks, decision trees, genetic algorithms, and non-parametric methods. In the proposed work a novel Feature Selection (FS) algorithm based on Ranker Search (RS) optimization method and Ensemble Genetic Search (EGS) is chosen for selecting the appropriate features and also class label refining for correcting misclassified instances from the dataset.

The anomaly detection methods for mobile ad hoc networks to detect the intrusions are used with genetic algorithm technique this audit data is reduced by means of feature selection technique. Intrusion Detection System (IDS) are software or hardware systems that automate the process of monitoring and analyzing the events that occur in a computer network, to detect malicious activity. The severity of attacks occurring in the network has increased drastically, Intrusion detection system have become a necessary addition to security infrastructure of most organizations.

Intrusion detection allows organization to protect their systems from the threats that come with increasing network connectivity and reliance on information systems. Intrusions are caused by: Attackers accessing the systems, Authorized users of the systems who attempt to gain additional privileges for which they are not authorized, Authorized users who misuse the privileges given to them. Various algorithms have been developed to identify different types of network intrusions. There is no heuristic to confirm the accuracy of their results. The exact effectiveness of a network intrusion detection system's ability to identify malicious sources cannot be reported unless a concise measurement of performance is available. Machine Learning includes a number of advanced statistical methods for handling regression and classification tasks with multiple dependent and independent variables. Methods include Support Vector Machines (SVM) for regression and classification, Naive Bayes for classification, and k-Nearest Neighbors (KNN) for regression and classification. Support Vector Machine (SVM) is one method that performs regression and classification tasks by constructing nonlinear decision boundaries. Support Vector Machines can exhibit a large degree of flexibility in handling classification and regression tasks of varied complexities Feature Extraction and Selection process involves the preprocess part, we use the packet sniffer, which is built with Jpcap library, to store network packet information including IP header, TCP header, UDP header, and ICMP header from each promiscuous packet.The packet information is divided by considering connections between any two IP addresses

(source IP and destination IP) and collects all records every 2 seconds.

Classification is a supervised machine learning procedure in which the effective model is constructed for prediction. The accuracy of classification mainly depends on the type of features and the characteristics of the dataset. Feature selection is an efficient approach in searching the most descriptive features which would contribute to the increase in the performance of the inductive algorithm by reducing dimensionality and processing time. In the present work a hybrid embedded feature selection algorithm with class label refining and handled numeric class problem in classifier are implemented. A novel feature selection algorithm based on ranker search optimization method and ensemble genetic search for selecting the appropriate features and class label refining for correcting misclassified instances from the dataset have been done. In general, techniques for dimensionality reduction focus either on selecting a proper subset from the original set of I attributes, or on mapping the initial I-dimensional data onto the K-dimensional space, where K<I. Many feature selection and extraction techniques can be found in the literature. Neural networks have been widely applied to a huge variety of supervised pattern classification problems. During the last decade, neural networks have also been successfully employed for feature selection and extraction The proposed system enhances a new feature selection strategy based on using neural networks. A three-layer multilayer perceptron (MLP) trained by the backpropagation algorithm is used as the tool to determine which attributes are to be removed from the original set.

2. RELATED WORKS

The work pertaining to classification of network packets includes C4.5, Support Vector Machine,CART – Classification and Regression Trees, Expectation and Maximisation Algorithm,K Nearest Neighbour etc. We can see the specifics cases in detail.

Given a set S of cases, C4.5 first grows an initial tree using the divide-and-conquer algorithm as follows:

If all the cases in S belong to the same class or S is small, the tree is a leaf labeled with the most frequent class in S. Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets S1, S2, . . . according to the outcome for each case, and apply the same procedure recursively to each subset.

Complex decision trees can be difficult to understand, for instance because information about one class is usually distributed throughout the tree. C4.5 introduced an alternative formalism consisting of a list of rules of the form “if A and B and C and ... then class X”, where rules for each class are grouped together. A case is classified by finding the first rule whose conditions are satisfied by the case; if no rule is satisfied, the case is assigned to a default class.

C4.5 rulesets are formed from the initial (unpruned) decision tree. Each path from the root of the tree to a leaf becomes a prototype rule whose conditions are the outcomes along the path and whose class is the label of the leaf. This rule is then simplified by determining the effect of discarding each condition in turn. Dropping a condition may increase the number N of cases covered by the rule, and also the number E of cases that do not belong to the class nominated by the rule, and may lower the pessimistic error rate determined as above. A hill-climbing algorithm is used to drop conditions until the lowest pessimistic error rate is found.

The k-means algorithm is a simple iterative method to partition a given dataset into a userspecified number of clusters, k. A detailed history of k-means along with descriptions of several variations . Gray and Neuhoff provide a nice historical background for k-means placed in the larger context of hill-climbing algorithms.The algorithm converges when the assignments (and hence the cj values) no longer change. Note that each iteration needs N × k comparisons, which determines the time complexity of one iteration. The number of iterations required for convergence varies and may depend on N, but as a first cut, this algorithm can be considered linear in the dataset size. One issue to resolve is how to quantify “closest” in the assignment step.

The default measure of closeness is the Euclidean distance, in which case one can readily show that the non-negative cost function, In today’s machine learning applications, support vector machines (SVM) are considered a must try—it offers one of the most robust and accurate methods among all well-known algorithms. It has a sound theoretical foundation, requires only a dozen examples for training, and is insensitive to the number of dimensions. In addition, efficient methods for training SVM are also being developed at a fast pace. In a two-class learning task, the aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data. The metric for the concept of the “best” classification function can be realized geometrically. For a linearly separable dataset, a linear classification function corresponds to a separating hyperplane f (x) that passes through the middle of the two classes, separating the two. Once this function is determined, new data instance xn can be classified by simply testing the sign of the function f (xn); xn belongs to the positive class if f (xn) > 0. Because there are many such linear hyperplanes, what SVM additionally guarantee is that the best such function is found by maximizing the margin between the two classes. Intuitively, the margin is defined as the amount of space, or separation between the two classes as defined by the hyperplane. Geometrically, the margin corresponds to the shortest distance between the closest data points to a point on the hyperplane. Having this geometric definition allows us to explore how to maximize the margin, so that even many of the pattern finding algorithms such as decision tree, classification rules and clustering techniques that are frequently used in data mining have been developed in machine learning research community.

Frequent pattern and association rule mining is one of the few exceptions to this tradition. The introduction of this technique boosted data mining research and its impact is tremendous. The algorithm is quite simple and easy to implement. Experimenting with Apriori-like algorithm is the first thing that data miners try to do. Finite mixture

distributions provide a flexible and mathematical-based approach to the modeling and clustering of data observed on random phenomena. We focus here on the use of normal mixture models, which can be used to cluster continuous data and to estimate the underlying density function. These mixture models can be fitted by maximum likelihood via the EM (Expectation–Maximization) algorithm.

PageRank was presented and published by Sergey Brin and Larry Page at the Seventh International World Wide Web Conference in April 1998. It is a search ranking algorithm using hyperlinks on the Web. Based on the algorithm, they built the search engine Google, which has been a huge success. Now, every search engine has its own hyperlink based ranking method. PageRank produces a static ranking of Web pages in the sense that a PageRank value is computed for each page off-line and it does not depend on search queries. The algorithm relies on the democratic nature of the Web by using its vast link structure as an indicator of an individual page's quality. In essence, PageRank interprets a hyperlink from page x to page y as a vote, by page x, for page y. One of the simplest, and rather trivial classifiers is the Rote classifier, which memorizes the entire training data and performs classification only if the attributes of the test object match one of the training examples exactly. An obvious drawback of this approach is that many test records will not be classified because they do not exactly match any of the training records. A more sophisticated approach, k-nearest neighbor (kNN) classification finds a group of k objects in the training set that are closest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighborhood. There are three key elements of this approach: a set of labeled objects, e.g., a set of stored records, a distance or similarity metric to compute distance between objects, and the value of k, the number of nearest neighbors. To classify an unlabeled object, the distance of this object to the labeled objects is computed, its k-nearest neighbors are identified, and the class labels of these nearest neighbors are then used to determine the class label of the object In addition, weights can be assigned to the training objects themselves. This can give more weight to highly reliable training objects, while reducing the impact of unreliable objects. The PEBLS system by Cost and Salzberg is a well known example of such an approach. KNN classifiers are lazy learners, that is, models are not built explicitly unlike eager learners (e.g., decision trees, SVM, etc.). Thus, building the model is cheap, but classifying unknown objects is relatively expensive since it requires the computation of the k-nearest neighbors of the object to be labeled. A number of techniques have been developed for efficient computation. Given a set of objects, each of which belongs to a known class, and each of which has a known vector of variables, our aim is to construct a rule which will allow us to assign future objects to a class, given only the vectors of variables describing the future objects. Problems of this kind, called problems of supervised classification, are ubiquitous, and many methods for constructing such rules have been developed. One very important one is the naive Bayes method—also called idiot's Bayes, simple Bayes, and independence Bayes. This method is important for several reasons. It is very easy to construct, not needing any complicated iterative parameter estimation schemes. This means it may be readily applied to huge data sets. It is easy to interpret, so users unskilled in classifier technology can understand why it is making the classification it makes. And finally, it often does surprisingly well: it may not be the best possible classifier in any particular application, but it can usually be relied on to be robust and to do quite well. General discussion of the naive Bayes method and its merits are given.

The 1984 monograph, "CART: Classification and Regression Trees," co-authored by Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone, represents a major milestone in the evolution of Artificial Intelligence, Machine Learning, non-parametric statistics, and data mining. The work is important for the comprehensiveness of its study of decision trees, the technical innovations it introduces, its sophisticated discussion of tree structured data analysis, and its authoritative treatment of large sample theory for trees. While CART citations can be found in almost any domain, farmore appear in fields such as electrical engineering, biology, medical research and financial topics than, for example, in marketing research or sociology where other tree methods are more popular. This section is intended to highlight key themes treated in the CART monograph so as to encourage readers to return to the original source for more detail. Missing values appear frequently in real world, and especially business-related databases, and the need to deal with them is a vexing challenge for all modelers. One of the major contributions of CART was to include a fully automated and highly effective mechanism for handling missing values. Decision trees require a missing value-handling mechanism at three levels: (a) during splitter evaluation,(b) when moving the training data through a node, and (c) when moving test data through a node for final class assignment. Regarding (a), the first version of CART evaluated each splitter strictly on its performance on the subset of data for which the splitter is available. Later versions offer a family of penalties that reduce the split improvement measure as a function of the degree of missingness. The CART mechanism discovers "surrogate" or substitute splitters for every node of the tree, whether missing values occur in the training data or not. The surrogates are thus available should the tree be applied to new data that does include missing values. This is in contrast to machines that can only learn about missing value handling from training data that include missing values. Friedman suggested moving instances with missing splitter attributes into both left and right child nodes and making a final class assignment by pooling all nodes in which an instance appears. Quinlan opted for a weighted variant of Friedman's approach in his study of alternative missing value-handling methods. Our own assessments of the effectiveness of CART surrogate performance in the presence of missing data are largely favorable, while Quinlan remains agnostic on the basis of the approximate surrogates he implements for test purposes. Friedman et al. noted that 50% of the CART code was devoted to missing value handling; it is thus unlikely that Quinlan's experimental version properly replicated the entire CART surrogate mechanism. Costs are central to statistical

decision theory but cost-sensitive learning received only modest attention before Domingos. Since then, several conferences have been devoted exclusively to this topic and a large number of research papers have appeared in the subsequent scientific literature. It is therefore useful to note that the CART monograph introduced two strategies for cost-sensitive learning and the entire mathematical machinery describing CART is cast in terms of the costs of misclassification. The cost of misclassifying an instance of class i as class j is C(i, j) and is assumed to be equal to 1 unless specified otherwise; C(i, i) = 0 for all i . The complete set of costs is represented in the matrix C containing a row and a column for each target class. Any classification tree can have a total cost computed for its terminal node assignments by summing costs over all misclassifications. The issue in cost-sensitive learning is to induce a tree that takes the costs into account during its growing and pruning phases.

3.PROPOSED SYSTEM

In fig 1 ,proposes a useful pre-processing step is to run your data through the following data cleaning routines. The proposed to enhance the Feature Selection process by using a new MLP based Feature Selection method. Supervised learning algorithms employ a collection of instances (or training set) to estimate the class label of new input samples. The instances are generally represented by a number of attributes (or features) and a class value. Not all of the attributes result equally important for a specific task. In fact, some features can be redundant or irrelevant.

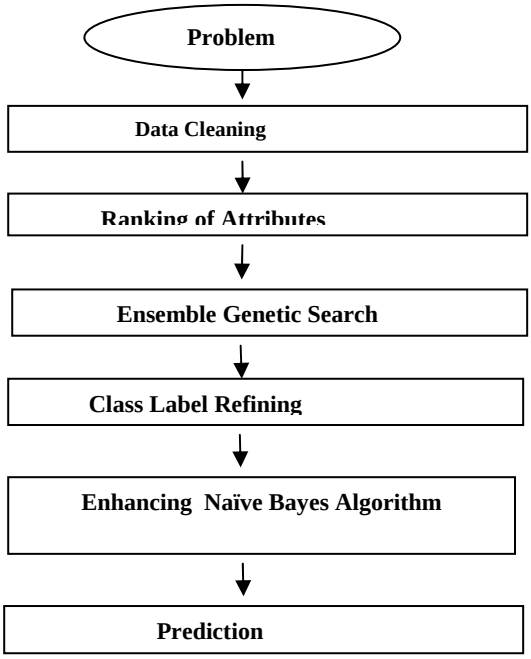

Figure 1. Proposed Architecture

We created a NIDS which will detect scan like port scan using SVM. It collects packet from the network for every 4 seconds. The change in frequency for normal packet and attack packet we train our SVM with normal and attack packets. When an unknown packet is coming SVM can classify easily whether it is a normal or attack packet. Using this method we could detect 95% of attack packets correctly and warning the administrator about it. According to the administrators decision log file is created and stored for future reference.

3.1 PRE-PROCESSING

We use the packet sniffer, which is built with Jpcap library, to store network packet information including IP header, TCP header, UDP header, and ICMP header from each promiscuous packet. After that, the packet information is divided by considering connections between any two IP addresses (source IP and destination IP) and collects all records every 2 seconds.

From real time network, we extracted 14 features.

No Feature Description Data Type
1 No: of TCP packets Integer
2 No: of TCP source port Integer
3 No: of TCP destination port Integer
4 No: of TCP fin flag Integer
5 No: of TCP syn flag Integer
6 No: of TCP reset flag Integer
7 No: of TCP push flag Integer
8 No: of TCP ack flag Integer
9 No: of TCP urget flag Integer
10 No: of UDP packets Integer
11 No: of UDP source port Integer
12 No: of UDP destination port Integer
13 No: of ICMP packets Integer
14 Answer class String (Normal, DoS, Probe)

Here feature selection is done using two different methods ROUGH SET and PCA. We did comparative study and selected rough set as optimal feature selection method.

3.2 ROUGH SET THEORY

Rough set theory can be regarded as a new mathematical tool for imperfect data analysis. The theory has found applications in many domains, such as decision support, engineering, environment, banking, medicine and others. Rough set philosophy is founded on the assumption that with every object of the universe of discourse some information (data, knowledge) is associated.

Feature selection is often isolated as a separate step in the processing of pattern sets. Features may be irrelevant (having no effect on the processing performance) or relevant (having an impact on the processing performance). A feature may have a different discriminatory or predictive power. We present rough sets methods and Principal Components Analysis in the context of feature selection in pattern classification. This section focuses the discussion on feature selection criteria including rough set-based methods.

3.3 ROUGH SETS AND FEATURE SELECTION

Rough sets theory has been proposed by Professor Pawlak for knowledge discovery in databases and experimental data sets. It is based on the concept of an upper and a lower approximation of a set, the approximation space and models of sets. Certain attributes in an information system may be redundant and can be eliminated without losing essential classificatory information. Rough sets provide a method to determine for a given information system the most important attributes from a classificatory power point of view. The concept of the reduct is fundamental for rough sets theory. A reduct is the essential part of an information

system (related to a subset of attributes) which can discern all objects discernible by the original set of attributes of an information system. The core and reduct are important concepts of rough sets theory that can be used for feature selection and data reduction.

3.4 ROUGH SETS IN KDD

Rough sets have many applications in KDD among them, feature selection, data reduction, and discretization. Rough sets can be used to find subsets of relevant (indispensable) features. It provides a mathematical tool that can be used to find out all possible feature subsets. In Feature selection problem, the purpose of using Rough sets is to find the indispensable features.

3.5.PCA BASED FEATURE SELECTION

Principal Component Analysis is a well-established technique for dimensionality reduction and multivariate analysis. Examples of its many applications include data compression, image processing, visualization, exploratory data analysis, pattern recognition, and time series prediction.lower dimensional vectors and then reconstructing the original set.

The model parameters can be computed directly from the data - for example by diagonalizing the sample covariance matrix. Compression and decompression are easy PCA summarizes the variation in correlated multivariate attributes to a set of non-correlated components, each of which is a particular linear combination of the original variables. The extracted non-correlated components are called Principal Components (PC) and are estimated from the eigenvectors of the covariance matrix of the original variables. The model which uses PCA detects and identifies intrusions by profiling normal network behaviour as well as various attack behaviours. This is very useful for preventing intrusions according to the associated individual type of attack. The model can also achieve real-time intrusion identification based on dimensionality reduction and on a simple classifier. In the proposed method, each network connection is transformed into an input data vector. PCA is employed to reduce the high dimensional data vectors and identification is thus handled in a low dimensional space with high efficiency and low usage of system resources. The distance between a vector and its reconstruction onto those reduced subspaces representing different types of attacks and normal activities is used for identification. The low computational expense of the distance allows a real-time performance of intrusion identification. without loss of generality. The goal is to separate the two classes by a function which is induced from available examples. The goal is to produce a classifier that will work well on unseen examples, i.e. it generalizes well.Many possible linear classifiers that can separate the data, but there is only one that maximizes the margin (maximizes the distance between it and the nearest data point of each class). This linear classifier is termed the optimal separating hyper plane. The IDS is designed to provide the basic detection techniques so as to secure the systems present in the networks that are directly or indirectly connected to the Internet Administrator to track down bad guys on the Internet whose very purpose is to bring your network to a breach point and make it vulnerable to attacks.

In the future, we will increase number of testing data for our system and to find vary of accuracy. We also hope to combine RST method and genetic algorithm to improve the accuracy of IDS.

4 .RESULTS AND DISCUSSIONS

The term SVM is typically used to describe classification with support vector methods and support vector regression is used to describe regression with support vector methods. SVM (Support Vector Machine) is a useful technique for data classification. The classification problem can be restricted to consideration of the two-class problem

The present system just displays the log information but doesn't employ any techniques to analyze the information present in the log records and extract knowledge. Thus comparing with SVM and RST,we got an accuracy of 95%.The system can be extended by incorporating Data Mining techniques to analyze the information in the log records which may help in efficient decision making. The present system only detects the attacks only the known attacks.

Table 1. Comparison Metrics

	SVM	**RoughSet**	**DataCleaning + Feature Selection by MLP +SVM**	**DataCleaning + Feature Selection by MLP + RoughSet**
No. of Features Selected	27	29	26	28
Accuracy	64.32%	72.65%	82.68%	94 .25%

4.2 SIMULATION RESULTS

By removing such redundant and irrelevant attributes, a classifier with higher predictive accuracy can often be obtained. The simulation results Fig 2,3&4 show better results and prediction accuracy

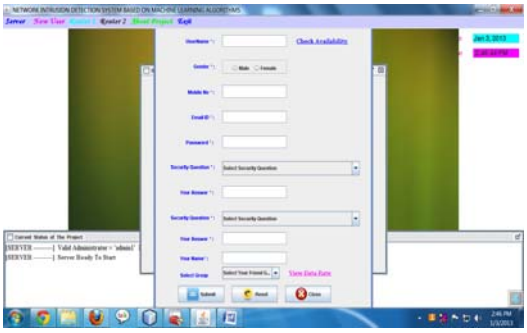

Figure 2. Simulation results

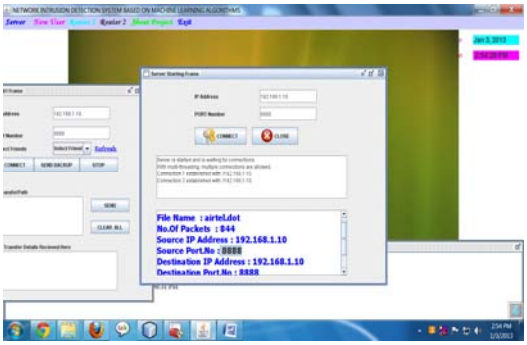

Figure 3. Simulation results

Figure 4. Simulation results

5.CONCLUSION

Classification is one of the important processes in data mining which is used to train and build a classifier or derive a set of rules based upon the given dataset a useful pre-processing step is to run your data through the following data cleaning routines. The proposed system is to enhance the Feature Selection process by using a new MLP based Feature Selection method. Supervised learning algorithms employ a collection of instances (or training set) to estimate the class label of new input samples. The instances are generally represented by a number of attributes (or features) and a class value. However, not all of the attributes result equally important for a specific task. In fact, some features can be redundant or irrelevant. By removing such redundant and irrelevant attributes, a classifier with higher predictive accuracy can often be obtained. The experimental results show better results and prediction accuracy.

REFERENCES

[1] Rung-Ching Chen, Kai-Fan Cheng and Chia-Fen Hsieh, "Using Rough Set and Support Vector Machine for Network Intrusion Detection", International Journal of Network Security & Its Applications (IJNSA), Vol 1, No 1, 2009.

[2] Phurivit Sangkatsanee, Naruemon Wattanapongsakorn and Chalermpol Charnsripinyo, "Real-time Intrusion Detection and Classification", IEEE network, 2009.

[3] Thomas Heyman, Bart De Win, Christophe Huygens, and Wouter Joosen, "Improving Intrusion Detection through Alert Verification", IEEE Transaction on Dependable and Secure Computing, 2004.

[4] Rung-Ching Chen, Kai-Fan Cheng and Chia-Fen Hsieh, "Using Rough Set and Support Vector Machine for Network Intrusion Detection", International Journal of Network Security & Its Applications (IJNSA), Vol 1, No 1, 2009.

[5] Cost S, Salzberg S (1993) A weighted nearest neighbor algorithm for learning with symbolic features. Mach Learn 10:57.78 (PEBLS: Parallel Examplar-Based Learning System).

[6] Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inform Theory 13(1):21–27.

[7] Dasarathy BV (ed) (1991) Nearest neighbor (NN) norms: NN pattern classification techniques. IEEE Computer Society Press.

[8] Zhang J, Kang D-K, Silvescu A, Honavar V (2006) Learning accurate and concise naïve Bayes classifiers from attribute value taxonomies and data. Knowl Inf Syst 9(2):157–179.

[9] Bezdek JC, Chuah SK, Leep D (1986) Generalized k-nearest neighbor rules. Fuzzy Sets Syst 18(3):237–256.

[10] VipinDas,Vijaya Pathak, Network Intrusion Detection system based on machine learnin alogorthms, International Journal of Computer Science & Information Technology (IJCSIT), Vol 2, No 6, December 2010