Comparative Analysis of Five Machine Learning Algorithms for IP Traffic Classification

Kuldeep Singh
University Institute of Engineering & Technology.
Panjab University
Chandigarh (India)
kuldeepsinghbrar87@gmail.com

Sunil Agrawal
University Institute of Engineering & Technology.
Panjab University
Chandigarh (India)
s.agrawal@hotmail.com

Abstract— **with rapid increase in internet traffic over last few years due to the use of variety of internet applications, the area of IP traffic classification becomes very significant from the point of view of various internet service providers and other governmental and private organizations. Now days, traditional IP traffic classification techniques such as port number based and payload based direct packet inspection techniques are seldom used because of use of dynamic port number instead of well-known port number in packet headers and various encryption techniques which inhibit inspection of packet payload. Current trends are use of machine learning (ML) techniques for this classification. In this research paper, real time internet traffic dataset has been developed using packet capturing tool and then using attribute selection algorithms, a reduced feature dataset has been developed. After that, five ML algorithms MLP, RBF, C4.5, Bayes Net and Naïve Bayes are used for IP traffic classification with these datasets. This experimental analysis shows that Bayes Net and C4.5 are effective ML techniques for IP traffic classification with accuracy in the range of 94 %.**

Keywords- IP Traffic Classification, Machine Learning, MLP, RBF, C4.5, Bayes Net, Naïve Bayes.

I. INTRODUCTION

In first decade of twenty first century, number of internet users increases very rapidly who use various internet applications in their day to day life. This leads to drastic increase in IP traffic. Various internet applications are www, e-mail, Web Media, FTP data, P2P, instant messaging, VoIP etc. IP Traffic Classification is an emerging issue for various internet service providers (ISPs) and other governmental and private organisations for various activities such as available bandwidth planning, fault diagnosis in network, analysis of quality of Service of any internet application or service, pricing information about users who use a particular internet application, Lawful Interception of internet traffic data by certain government agencies for various security related issues [1], [2].

The major contribution in internet traffic is done by peer to peer (P2P) applications such as Bit Torrent, Emule, Kaaza etc which leads to 80% rise in internet traffic [2]. Now a days, various multimedia websites such as Youtube also contribute in internet traffic to large extent. Common internet applications such www, e-mails and file transfer websites etc also have a significant amount of share in this traffic. Most of the users in peak traffic hours use various messenger based applications such as Yahoo Messenger, Google Talk for audio and video calls and for instant messaging which is again a major reason to rise in IP traffic.

Tradition IP traffic classification techniques are direct packet inspection based techniques such as port number based and payload based techniques [1], [3]. In present time, these techniques are rarely used. Port number based techniques become ineffective because of use of dynamic port number instead of well-known port number in packet headers. While payload based technique are also seldom used because privacy policies of government and certain cryptographic techniques which are used to encrypt packet payload, inhibit inspection of IP traffic packets.

In current scenario, Machine Learning (ML) techniques [1], [2], [10] are utilised for IP traffic classification which are based on training a network using a set of previous examples and then using that trained network for predicting classes of unknown test samples. In this paper, real time internet traffic dataset has been developed and then using attribute selection algorithm, a reduced feature dataset has also been developed. Then using this full feature and reduced feature datasets, five ML algorithms have been used for IP traffic classification: Multilayer Perceptron (MLP), Radial Basis Function Neural Network (RBF), C 4.5 Decision Tree Algorithm, Bayes Net Algorithm and Naïve Bayes Algorithm [10]. Performance of all these classifiers is analysed on the basis of classification accuracy, training time of classifiers, recall and precision values of classifiers for individual internet applications [1], [5].

The remaining paper is organised as follows: section II gives some information about related work done by various researchers in the field of IP traffic classification. Section III includes introductory information about all classifiers mentioned above. Section IV gives overview of internet traffic dataset. Implementation and result analysis is given in section V. Section VI includes conclusions and future scope.

978-1-4577-0240-2/11/$26.00 ©2011 IEEE

II. RELATED WORK

IP traffic classification is an emerging field in which various researchers have shown their interest over last few years. We have reviewed their papers for this research work. Some previous work done in this field by some researchers is discussed as follows:

In [1], Nguyen et al. have given a brief introduction to various machine learning techniques for IP traffic classification. They have discussed port number based and payload based IP classification techniques. Various machine learning techniques based upon clustering, supervised learning and hybrid approaches are explained briefly. Under clustering approach flow clustering using expectation maximization, automated application identification using Auto Class, TCP-based application identification using K- means, Identification of web and peer to peer in the network core techniques are explained. Under supervised learning approach, Statistical signature based technique using NN, LDA and QDA algorithms, Bayesian analysis techniques for, Real-time traffic classification using Multiple sub flows features, Generic algorithm based classification techniques etc are explained.

In [3], Runyuan Sun et al. have designed host based traffic collection platform to collect internet traffic of web, P2P and other applications. They have used three techniques for IP traffic classification such as Probabilistic neural network (PNN), RBF neural network and Support Vector Machine (SVM). They have concluded that PNN gives better performance as compared to other two networks. But their scope of research is limited to web and P2P applications only because they have not taken into account various other internet applications.

In [4], we have performed IP classification using RBF neural network and Back Propagation neural network. Performance of these networks is analyzed on the basis of classification accuracy, recall of individual applications, training time of networks and number of hidden layer neurons of the networks. In this work, it is concluded that RBF neural network gives very good performance as compared to back propagation neural network at the cost of very high training time and computational complexity because at 1000 hidden layer neurons, RBF network gives 90.10 % classification accuracy. But training time is 432 minutes. Therefore, this approach is not suitable for online traffic classification. There is still scope of further improvement in performance by using other ML algorithms for internet traffic classification.

III. VARIOUS MACHINE LEARNING CLASSIFIERS

In this paper, five famous machine learning algorithms are used which are explained in brief as follows:

A. Multilayer Perceptron

Multilayer Perceptron (MLP), [7], [8], [14] also known as Back Propagation Neural Network, is a feed forward multilayer artificial neural network which is based upon extended gradient-descent based Delta learning rule, commonly known as Back Propagation rule. In this network, error signal between desired output and actual output is being propagated in backward direction from output to hidden layer and then to input layer in order to train the network.

Consider the network shown in fig. 1. It consists of input layer having i neurons, hidden layer having j neurons and output layer having k neurons.

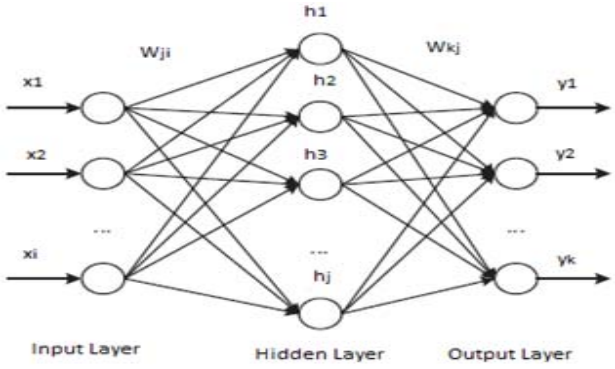

Figure 1. Multilayer Perceptron

In this research work, single hidden layer MLP is being used for IP traffic classification with learning rate of 0.3 and momentum term of 0.2 [11].

B. Radial Basis Function Neural Network

Radial Basis Function (RBF) Neural Network [4], [7], [14] is a multilayer feed forward artificial neural network in which radial basis functions are used as activation functions at each hidden layer neurons. The output of this RBF neural network is weighted linear superposition of all these basis functions.

The basic structure of RBF neural network is shown in fig 2. In this network, weights are fixed for input-hidden layer interconnections. While the weights for hidden-output layer interconnections are trainable. Each hidden layer neuron have a basis function $f_m(.)$.For any input vector X, the output of this network is given by following interpolation function as:

$$Y(X) = \sum_{m=0}^{M} W_i f(||X - X_i||) \quad (1)$$

Where $f(||X - X_i||)$ are M basis functions consisting of Euclidean distance between applied input X and training data point Xi.

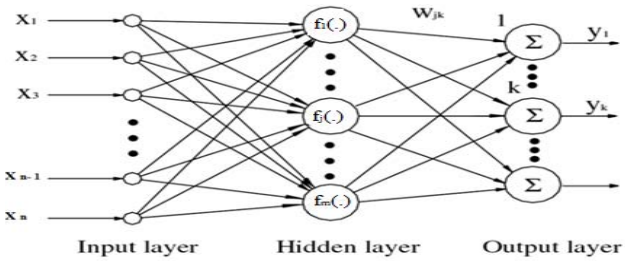

Figure 2. Radial Basis Function Neural Network

The commonly used basis function in RBF Algorithm is Gaussian function which is given as follows:

$$f(X) = \exp\left(-\frac{\|X-\mu\|}{2\sigma^2}\right) \quad (2)$$

Where μ is the Center and σ is spread constant which have direct effect on the smoothness of interpolating mapping function Y(X).

In this research work, single hidden layer RBF algorithm has been used for IP traffic classification with number of center points in hidden layer equal to 5 and 2 for full feature dataset and reduced feature dataset respectively [11].

C. C4.5 Algorithm

C4.5 is an ML algorithm used to generate Univariate decision tree [9]. It is an extension of Iterative Dichotomiser 3 (ID3) algorithm which finds simple decision trees. C4.5 is also called Statistical Classifier because its decision trees can be used for classification.

C4.5 builds decision trees from a set of training data in similar manner as ID3, using the concept of information entropy. The training dataset consists of various training samples which are characterized by multiple features and also consists of target class.

At each node of the tree, C4.5 chooses one feature of the data that splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain that is obtained from choosing a feature for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. After that, the C4.5 algorithm recurs on the smaller sub lists.

In this research work, C4.5 algorithm has been used for IP traffic classification with confidence factor of 0.25, minimum no. of instances per leaf equal to 2, no. of folds for pruning equal to 3 and seed used for randomizing the data, when error reduced pruning is used, equal to 1[11].

D. Bayes Net

Bays' Net (Bayesian Network), [10], [12] also known as Belief Network, is a probabilistic graphical model. This graphical structure is used to represent knowledge about an uncertain domain. In this model, each node represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables. These conditional dependencies in the graph are often estimated by using known statistical and computational methods.

Learning process in Bayesian Network take place in two steps: first learn a network structure, then learn the probability tables.

There are various approaches used for structure learning and in Weka tool, the following approaches are mainly considered:

- Local score metrics
- Conditional independence test
- Global score metrics
- Fixed structure

For each of these approaches, different search algorithms are implemented in Weka, such as hill climbing, simulated annealing and tabu search. Once a good network structure is identified, the conditional probability tables for each of the variables can be estimated.

In this research work, Bays' Net algorithm with simple estimator and K2 search algorithm has been used for IP traffic classification [10], [11].

E. Naïve Bayes

A Naïve-Bays' (NB) ML algorithm [12], [13] is a simple structure that has the class node as the parent node of all other nodes. Fig. 3 shows a basic structure of Naïve Bayes Classifier in which C represents main class and a, b, c and d represents other feature or attribute nodes of a particular sample. No other connections are allowed in a Naïve-Bayes structure. Naïve-Bayes has been used as an effective classifier. It is easy to construct Naïve Bayes classifier as compared to other classifiers because the structure is given *a priori* and hence no *structure* learning procedure is required. Naïve-Bayes assumes that all the features are independent of each other.

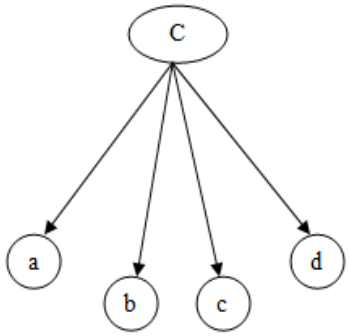

Figure 3. Naïve Bayes Classifier

Naïve-Bays' works very well over a large number of datasets, especially where the features used to characterize each sample are not properly correlated.

IV. INTERNET TRAFFIC DATASET

In this research work, a packet capturing tool, Wire shark, [15] is used to capture real time internet traffic. Wire shark is a network packet analyzer which captures network packets and displays that packet data as detailed as possible. Real time internet traffic packets are captured for the duration of 2 minutes for each individual application just considering on-going middle session of each application. During this packet capturing process, starting and end of each application are not taken into account.

In this process of developing datasets, two datasets are obtained: one is full feature dataset and another is reduced feature dataset [6]. In both the datasets, seven internet applications are taken into account such as www, e-mail, web media, P2P, FTP data, instant messaging and VoIP. These datasets include 2800 samples.

In full feature dataset, each data sample is characterized by 261 features which mainly consist of minimum, maximum, mean, variance and total values of no. of packets, average packets per second, packet size, duration, no. of conversations etc for Ethernet, IPv4, IPv6, TCP and UDP packets. While reduced feature dataset is obtained from full feature dataset using cfsSubsetEval evaluator and Best First search in attribute selection filter of Weka tool [11].

For our work, we have used 2.27 GHz Intel core i3 CPU workstation with 3GB of RAM and Microsoft Windows 7 operating system.

V. Implementation And Result Analysis

A. Methodology

In this research work, Weka tool, [11] which is a well-known data mining tool, is used for implementing IP traffic classification with five different ML algorithms. Two different internet traffic datasets namely, full feature dataset and reduced feature dataset, consisting of 2800 data samples in each, are divided into two sets consisting of 2500 data samples for training and 300 data samples for testing purpose for both datasets.

In this work, classification accuracy, training time, recall and precision values [1], [4] of individual internet application samples are employed in order to evaluate performance of these five ML algorithms/classifiers. All these parameters are defined as follows:

- Classification Accuracy: It is the percentage of correctly classified samples over all classified samples.
- Training Time: It is the total time taken for training of a machine learning classifier. In this paper, it is measured in seconds.
- Recall: It is the proportion of samples of a particular class Z correctly classified as belonging to that class Z. It is equivalent to True Positive Rate (TPR). In this paper, its value ranges from 0 to 1.
- Precision: It is the proportion of the samples which truly have class z among all those which were classified as class z. In paper its value ranges from 0 to 1.

B. Results and Analysis

Table I shows classification accuracy and training time of MLP, RBF, C4.5 Bayes Net and Naïve Bayes ML classifiers for full feature dataset.

It is clear from this table and figure 4 that maximum classification accuracy is provided by Bayes Net classifier for full feature dataset which is 85.33 %. From table I, it is evident that training time of Bayes Net classifier is 14 second which is much lesser as compared to that of MLP and RBF classifiers in case of full feature dataset. But it is slightly larger than that of C4.5 and Naïve Bayes classifiers.

TABLE I. CLASSIFICATION ACCURACY AND TRAINING TIME OF FIVE ML CLASSIFIERS FOR FULL FEATURE DATASET

ML Classifiers	MLP	RBF	C4.5	Bayes Net	Naïve Bayes
Classification Accuracy (%)	31.33	82.33	79	85.33	68
Training Time (Seconds)	220	126	12	14	4

TABLE II. CLASSIFICATION ACCURACY AND TRAINING TIME OF FIVE ML CLASSIFIERS FOR REDUCED FEATURE DATASET

ML Classifiers	MLP	RBF	C4.5	Bayes Net	Naïve Bayes
Classification Accuracy (%)	43	77	93.66	90	74.66
Training Time (Seconds)	56	89	1	2	1

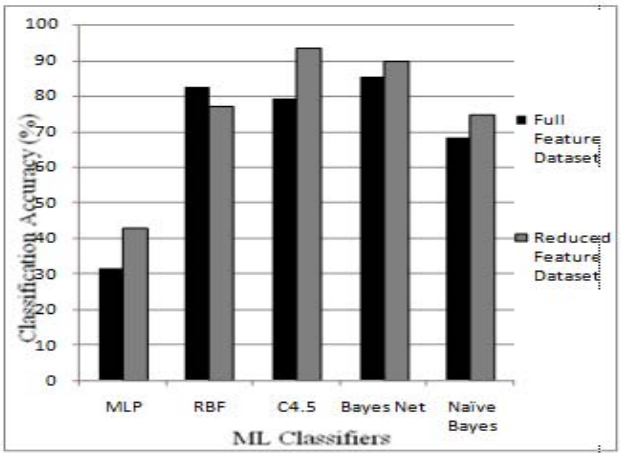

Figure 4. Classification Accuracy of five ML Classifiers for Full Feature and Reduced Feature Dataset

From these results, it is evident that Bays' Net gives better performance in terms of classification accuracy as compared to other four ML classifiers for full feature dataset. Figure 5 and 6 shows recall and precision values of three most accurate ML classifiers i.e. RBF, C4.5 and Bayes Net for individual internet applications.

Bays' Net gives 100% recall value for P2P, FTP data, instant messaging and VoIP applications. Similarly, it gives 100 % precision for Web media, P2P, IM and VoIP applications.

Thus it is again clear that Bays' Net gives better performance in terms of Recall and precision for most of internet applications in case of full feature dataset.

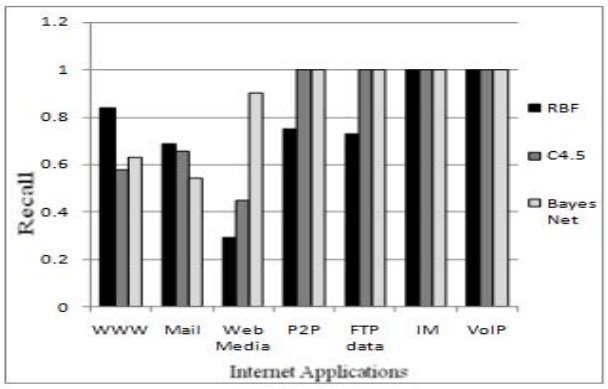

Figure 5. Recall of three most accurate ML Classifiers for Full Feature Dataset

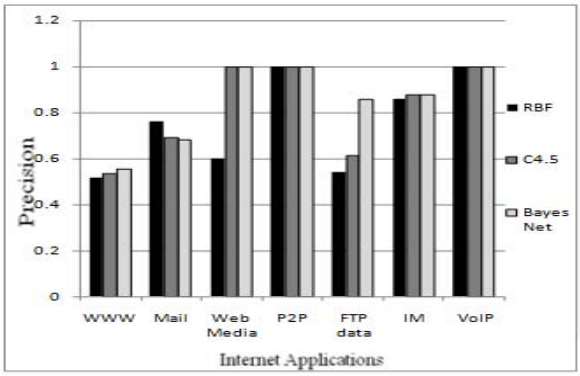

Figure 6. Precision of three most accurate ML Classifiers for Full Feature Dataset

Although Bays' Net provides better classification performance at full feature dataset. But there is still scope of further improvement in classification accuracy and reduction in training time and computational complexity if no. of features used to characterize each internet application can be reduced to great extent. Therefore, reduced feature dataset is also taken into account for performance analysis of these five ML classifiers.

Table II shows classification accuracy and training time of these five ML classifiers for reduced feature dataset. It is clear from this table and figure 4 that there is large degree of improvement in classification accuracy of all these ML classifiers. But maximum classification accuracy is provided by C4.5 classifier for reduced feature dataset which is 93.66 %. From table II, it is evident that training time of these classifiers is reduced to much extent. But training of most accurate ML classifier i.e. C4.5 Classifier is only 1 second which is much lesser as compared to that of MLP, RBF and Bays' Net classifiers in case of reduced feature dataset.

From these results, it is clear that C4.5 gives better performance in terms of classification accuracy and training time as compared to other four ML classifiers for reduced feature dataset. But classification accuracy of Bays' Net is also very high i.e. 90% in case of reduced feature dataset. Figure 7 and 8 shows recall and precision values of three most accurate ML classifiers i.e. RBF, C4.5 and Bays' Net for individual internet applications. C4.5 gives 100% recall value for P2P, FTP data, instant messaging and VoIP applications. Similarly, it gives 100 % precision for Web media, P2P and VoIP applications. Thus It is again very clear that C4.5 gives better performance in terms of Recall and precision for most of internet applications in case of reduced feature dataset.

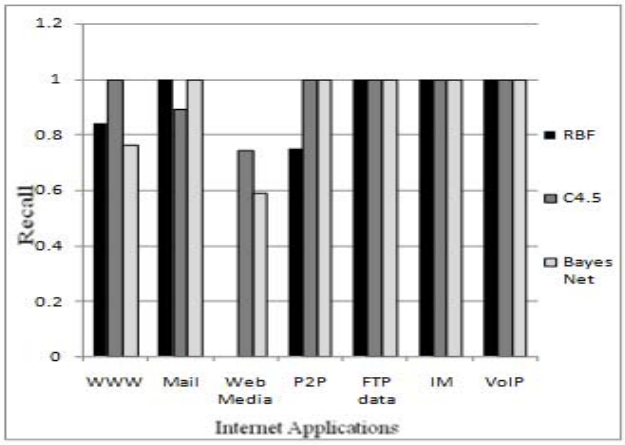

Figure 7. Recall of three most accurate ML Classifiers for Reduced Feature Dataset

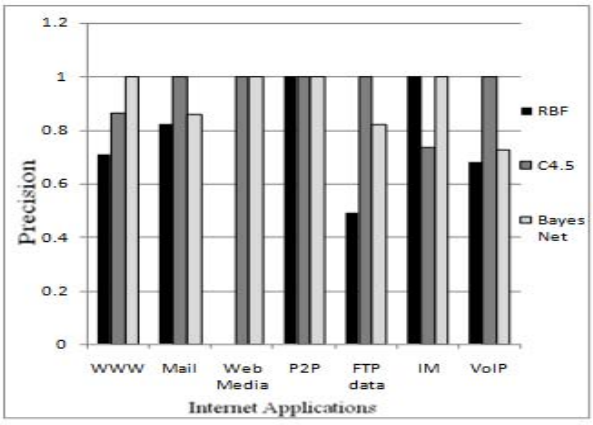

Figure 8. Precision of three most accurate ML Classifiers for Reduced Feature Dataset

From this analysis, it is evident that Bayes Net gives good performance in terms of classification accuracy, training tine, recall and precision of individual internet applications for full feature dataset. But this performance is still not remarkable. So by reducing the number of features used to characterize internet applications, classification performance is further improved. In this case, C4.5 gives highest degree of performance in terms of all the factors mentioned above.

VI. Conclusions And Future Scope

In this paper, firstly real time internet traffic has been captured using Wire shark software which is a packet capturing tool. After that, Internet traffic is classified using five ML classifiers. Results show that Bays' Net gives better classification of internet traffic data in terms of classification

accuracy, training time of classifiers, recall and precision values of classifiers for individual internet applications. After that, the no. of features used to characterize each internet application data sample of this dataset are further reduced to make a reduced feature dataset. Results show that with reduced feature dataset, performance of these classifiers is improved to large extent. In this case, C4.5 classifier gives very much accurate results. Thus it is evident that Bays' Net and C4.5 are effective ML techniques for IP traffic classification with accuracy in the range of 94 %.

In this research work, internet traffic dataset has been developed by considering packet flow duration of 2 minutes for each application which is still very large. This flow duration can be further reduced to make analysis more real time compatible. Secondly, internet traffic can also be captured from various different real time environments such as university or college campus, offices, home environments etc. This internet traffic dataset can be extended for many other internet applications.

REFRENCES

[1] Thuy T.T. Nguyen and Grenville Armitage. "A Survey of Techniques for Internet Traffic Classification using Machine Learning," *IEEE Communications Survey & tutorials,* Vol. 10, No. 4, pp. 56-76, Fourth Quarter 2008.

[2] Arthur Callado, Carlos Kamienski, Géza Szabó, Balázs Péter Ger˝o, Judith Kelner,Stênio Fernandes ,and Djamel Sadok. "A Survey on Internet Traffic Identification," *IEEE Communications Survey & tutorials,* Vol. 11, No. 3, pp. 37-52, Third Quarter 2009.

[3] Runyuan Sun, Bo Yang, Lizhi Peng, Zhenxiang Chen, Lei Zhang, and Shan Jing. "Traffic Classification Using Probabilistic Neural Network," in *Sixth International Conference on Natural Computation (ICNC 2010*), 2010, pp. 1914-1919.

[4] Kuldeep Singh and Sunil Agrawal, "Internet Traffic Classification using RBF Neural Network," in *International Conference on Communication and Computing technologies(ICCCT-2011)*, Jalandhar, India, February 25-26, 2011, paper 10, p.39-43.

[5] Luca Salgarelli, Francesco Gringoli, Thomas Karagiannis. "Comparing Traffic Classifiers*," ACM SIGCOMM Computer Communication Review*, Vol. 37, No. 3, pp. 65-68, July 2007.

[6] Andrew W. Moore, Denis Zuev, Michael L. Crogan, "Discriminators for use in flow-based classification," *Queen Mary University of London, Department of Computer Science,* RR-05-13, August 2005.

[7] Y.L. Chongand K. Sundaraj, "A Study of Back Propagation and Radial Basis Neural Networks on ECG signal classification*," in 6th International Symposium on Mechatronics and its Applications (ISMA09),* Sharjah, UAE*, March 24-26, 2009.*

[8] Mutasem khalil Alsmadi, Khairuddin Bin Omar, Shahrul Azman Noah ,Ibrahim Almarashdah, "Performance Comparison of Multi-layer Perceptron (Back Propagation, Delta Rule and Perceptron) algorithms in Neural Networks" in *2009 IEEE International Advance Computing Conference (IACC 2009) ,*Patiala, India*,* 6-7 March 2009, p. 296-299.

[9] Thales Sehn Korting, "C4.5 algorithm and Multivariate Decision Trees," Image Processing Division, National Institute for Space Research – INPE, SP, Brazil.

[10] Ian H. Witten and Eibe Frank*, Data Mining: Practical Machine Learning Tools and Techniques*, 2th edition, Morgan Kaufmann Publishers, San Francisco, CA, 2005.

[11] Weka website. Available: http:// www.cs.waikato.ac.nz/ml/weka/

[12] Jie Cheng, Russell Greiner, "Learning Bayesian Belief Network Classifiers: Algorithms and System," *Department of Computing Science, University of Alberta,* Edmonton, Alberta, Canada.

[13] Ioan Pop, "An approach of the Naive Bayes classifier for the document classification," *General Mathematics,* Vol. 14, No. 4, pp.135-138, 2006.

[14] Simon Hakin, *Neural Networks: A Comprehensive foundation,* 2th edition, Pearson Prentice Hall, New Delhi, 2005.

[15] Wireshark, Available: http:// www.wireshark.org

.