Machine Learning Classifier for Internet Traffic from Academic Perspective

S. Agrawal
Panjab University
Chandigarh, India

Jaspreet Kaur
Panjab University
Chandigarh, India

B.S.Sohi
Campus Director, CGC
Gharuan, Punjab, India

ABSTRACT

The infinite number of websites in the internet world can be classified into different categories in different ways. But if we talk about the educational institutions, websites can be classified into two categories, educational websites and non-educational websites. Educational websites are those websites which are used by the students to acquire knowledge, to explore educational topics, for the research work etc. The non-educational websites are used for entertainment and to keep in touch with people and to get to know more people. In educational institutes for the optimum use of network resources and for the welfare of the students, the use of non-educational websites should be banned while only the educational websites should be allowed to access. Recent trends are use of ML (machine learning) algorithms for internet traffic classification. In this paper, we use three ML classifiers Bayes Net, C4.5 and Radial basis function (RBF) neural network to classify the educational and non-educational websites and compare their performances. Results show that Bayes Net gives best performance for intended classification of internet traffic in terms of classification accuracy, training time of classifiers, recall and precision.

General Terms

Educational websites, Non-educational websites, Internet traffic classification, Machine learning.

Keywords

RBF, C4.5, Bayes Net, features.

1. INTRODUCTION

In the recent years with the rapid growth in internet users, the internet traffic is going to be increased at drastic rate both for educational and non-educational purposes. As the use of a number of internet applications by users in different field increases, the internet traffic also increases day by day.

Internet service providers as well as enterprise networks require the ability to accurately identify the different applications, for a range of uses, including network operations and management, application-specific traffic engineering, capacity planning, resource provisioning, service differentiation and cost reduction.

There is infinite number of websites in this world of the internet. There may be different ways to classify these websites depending on the motivation for classification. Like one can classify them from academic perspectives, as educational and non-educational websites. Educational websites are used for educational purposes that are to acquire knowledge in any educational field. Similarly non-educational websites can be used for entertainment and to keep in touch with people and to get to know more people.

There are a number of Educational websites e.g. www.ieeeexpore.ieee.org IEEE is the world’s largest professional association dedicated to advancing technological innovation and excellence for the benefit of humanity. IEEE and its members inspire a global community through IEEE's highly cited publications, conferences, technology standards, and professional and educational activities. www.sciencedirect.com, according to Wikipedia (January 2012) Science Direct is a leading full-text scientific database offering journal articles and book chapters from more than 2,500 peer-reviewed journals and more than 11,000 books. There are currently more than 9.5 million articles/chapters, a content base that is growing at a rate of almost 0.5 million additions per year, www.math.com is used for solving mathematical problems, www.novelguide.com is used for literary analysis, www.sparknotes.com is used for study guides for literature, poetry, history, film and philosophy etc. Non educational websites like www.bittorrent.com, www.yahoomassenger.com and www.movies.com etc. websites used for chatting purposes and for songs, movies and games download etc. also come under the category of non-educational websites.

Social networking or non-educational surfing is a recent invention that has the Internet still at the edge of its seat due to its popularity with people. This is mostly because it really is for the people. Bringing every kind of social group together in one place and letting them interact is really a big thing indeed. Although there are advantages of social websites like Low Costs, Builds Credibility, Connections. But there are more dominating disadvantages like Lack of Anonymity, Scams and Harassment, Time Consuming etc.

Moreover, from Wikipedia, Orkut is one of the most visited websites in India and Brazil. As of October 2011, 59.1% of orkut's users are from Brazil, followed by India with 27.1% and Japan with 6.7%. As of January 2012, Facebook has more than 800 million active users. Also, based on ConsumersReports.org in May 2011, there are 7.5 million children under 13 with accounts, violating the site's terms of service. The major contribution in internet traffic is done by peer to peer (P2P) applications such as Bit Torrent, Emule, Kaaza etc. leads 80% rise in internet traffic [1]. So the surfing of non-educational websites increases more rapidly as compared to educational websites.

Internet Traffic classification can be done either offline or online. In online classification, analysis is performed while data packets flowing through the network are captured; but in case of offline classification technique, firstly data traces are captured and stored and then analyzed/classified later [1].

Traditional IP traffic classification techniques are direct packet inspection based techniques such as port number based and payload based techniques [2], [3]. But presently these techniques are rarely used because of their inherent limitations. In payload based technique , payload of few TCP/IP packets are analyzed in order to identify any particular application which is not possible today because of use of cryptographic techniques used to encrypt data in packet payload and privacy policies of governments which do not allow any unaffiliated third party to inspect each packets payload.

In port number based packet inspection technique, well-known port numbers are provided in header of IP packets which are reserved by IANA (Internet Assigned Numbers Authority) for particular applications e.g. port number 80 is reserved for web based applications [4]. Unfortunately, this method is also rarely used due to the use of Dynamic port numbers instead of Well-known port numbers for various applications.

Now numbers of researchers are looking for internet traffic classification techniques without deep inspection of packets. Most of these techniques come under the category of Machine Learning (ML) techniques. In ML techniques, first, features are defined to identify and differentiate future unknown internet traffic data. These features are attributes of flows calculated over multiple packets (such as maximum or minimum packet lengths in each direction, flow durations or inter-packet arrival times, data rate of traffic, traffic volume etc.) [5].

In our research work, internet traffic is to be classified into two classes, one for educational websites and another for non-educational websites. There are infinite websites in educational institutes like educational and non-educational, that one can access, but for the optimum use of network resources in the educational institutes, the use of non-educational websites should be banned while only the educational websites should be allowed to open.

The remaining paper is organised as follows: section 2 gives some information about related work done by various researchers in the field of IP traffic classification. Section 3 includes ML algorithms, section 4 gives dataset creation. Section 5 gives methodology and result analysis and section 6 gives conclusion.

2. RELATED WORK

IP traffic classification is an emerging field which is the choice of number of researchers over last couple of years. Because of limitations of port number based and payload based techniques, researchers are looking at ML techniques. For this research work, numbers of research papers have been reviewed. Some of the previous work done in this field by some researchers is discussed as follows:

In [6, 7], Zander et al. have discussed their work by looking at maximizing intra-cluster homogeneity (or cluster purity) by investigating, which set of features separate the flows from different applications with greatest accuracy. The traces used in this analysis are from a publicly available archive of traces and port-based analysis was used to establish the “base truth". The authors continued this work and recently used the C4.5 supervised machine learning algorithm to estimate the traffic trends in archival traces.

In [8], Soysal and Schmidt have presented a systematic approach for investigating and evaluating the internet traffic classification performance of three supervised Machine Learning (ML) algorithms namely Bayesian Networks (BNs), Decision Trees (DTs) and Multilayer Perceptrons (MLPs), using flow traces. The performance results indicate that DTs have both a higher accuracy and a higher classification rate than BNs. However, DTs require a larger build time and are more susceptible in the case of incorrect or small amounts of training data. A detailed analysis of traffic classification with MLPs that are trained by back propagation is carried out to identify the drawbacks of this algorithm. As a result, it is not possible to simultaneously achieve acceptable recall values for these traffic types when the MLP algorithm is used.

In [9], Arya and Mishra have proposed multilevel classifiers based on the performance of multiple classifiers for internet traffic classification. Five classifiers namely J48, Random Tree, Random Forest, Bagging and boosting algorithms are evaluated over single benchmark dataset. Proposed multilevel classifiers give better performance than single classifier. Performance of classifier for P2P class increases by using classifier combinations using Bagging and Multiboosting. Multiboosting outperforms the Bagging approach.

In [10], Kuldeep Singh and Agrawal have performed IP traffic classification using RBF neural network and Back Propagation neural network. This paper concludes that RBF neural network gives better performance as compared to back propagation neural network. But training time and computational complexity of RBF network is extremely high. At 1000 hidden layer neurons, RBF network gives 90.10 % classification accuracy. But training time is 432 minutes. Therefore, this technique is not effective for online IP traffic classification. Better classification performance can be obtained by using other ML techniques.

In [11] Youngli Ma et al. integrated theory with actual needs on the measure works, and made use of the characteristics of the network traffic that were understood easily in internet network. First, they used the CFS and genetic search method to select three subsets from three full sets. Then they primarily selected 15 kinds of algorithms from more than 50 ones which involved in decision tree, rules, Bayes, neural network algorithms, finally proposed the multivariate evaluation method (MEM) to assess these algorithms on accuracy, memory consumption, CPU utilization, construction model time and test time.

In [12] Singh and Agrawal captured firstly real time internet traffic 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. Their 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 %.

3. MACHINE LEARNING ALGORITHMS

In this research paper, three well-known machine learning algorithms are used which are reported in different research papers to be performing well in most of the applications. These machine learning algorithms are discussed in brief as follows:

3.1 Radial Basis Function Neural Network

Radial Basis Function (RBF) Neural Network [10, 13, 14] is a multilayer feed forward artificial neural network which uses radial basis functions as activation functions at each hidden layer neuron. The output of this RBF neural network is weighted linear superposition of all these basis functions.

The basic model of RBF neural network is shown in figure 1. In this network, weights for input-hidden layer interconnections are fixed, while the weights are trainable for hidden-output layer interconnections. Each neuron in hidden layer has basis function $U_m(.)$. For any input vector X, the output of this network is given by following input - output mapping function as:

$$\mathrm{Y(X)} = \sum_{i=0}^{m} \mathrm{Wi}\ U(\,||\mathrm{X} - \mathrm{Xi}||) \qquad (1)$$

Where $\mathrm{U}(||\mathrm{X} - \mathrm{Xi}||)$ is M basis functions consisting of Euclidean distance between applied input X and training data point Xi.

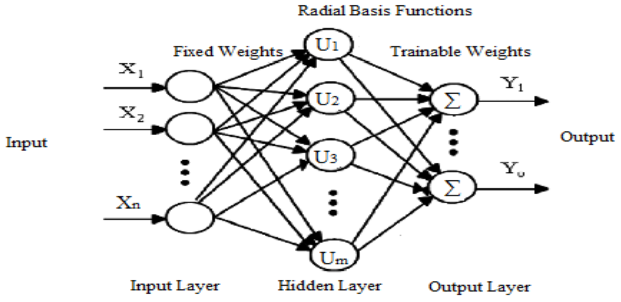

Figure 1. Radial Basis Function Neural Network

The commonly used basis function in RBF Algorithm is Gaussian basis function which is given by following formula:

$$\mathrm{U(X)} = \exp\left(-\frac{||\mathrm{X}-\mu||}{2\sigma^2}\right) \qquad (2)$$

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

3.2 C4.5 Algorithm

C4.5 is a well-known decision tree Machine Learning algorithm used to generate Univariate decision tree [15]. It is an extension of Iterative Dichotomiser 3 (ID3) algorithm which is used to find simple decision trees. C4.5 is also called as Statistical Classifier due of its classification capability.

C4.5 makes decision trees from a set of training data samples, with the help of information entropy concept. The training dataset consists of large number of training samples which are characterized by various features and it also consists of target class. C4.5 selects one particular feature of the data at each node of the tree which is used to split its set of samples into subsets enriched in one or another class. It is based upon the criterion of normalized information gain that is obtained from selecting a feature for splitting the data. The feature with the highest normalized information gain is selected and a decision is made. After that, the C4.5 algorithm repeats the same action on the smaller subsets.

In present 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 for dataset [16].

3.3 Bayes Net Algorithm

Bayes Net (Bayesian Network), [17, 18] popularly known as Belief Network, is a probabilistic graphical model. This graphical model 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 those corresponding random variables. These conditional dependencies in the graph are estimated by using known statistical and computational methods.

Learning of Bayesian Network takes place in two phases: first learning of a network structure and then learn the probability tables. There are various approaches used for structure learning and in Weka tool, the following approaches are mainly taken into account:

- 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 present work, Bayes Net algorithm with simple estimator and K2 search algorithm has been used for IP traffic classification [17, 16].

4. DATASET

In this research work, Wireshark, [19], which is well-known open-source packet capturing software, is used to capture internet traffic related to educational and non-educational internet applications. It is a network packet analyzer which is used to capture network packets and extract detail of the captured packet. To create data set, Internet traffic packets are captured for the duration of 1 minute for each educational and non-educational website by considering on-going middle session as well as starting and end of each application. 108 features are extracted for each website out of which, six features are extracted directly from statistics summary of Wireshark. While other 102 features are extracted for TCP and UDP conversations of Wireshark.

In this process of packet capturing and feature extraction, a dataset of 1740 samples is developed by performing feature extraction of traffic traces using MATLAB code. In this dataset, each sample is characterized by 108 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 TCP and UDP packets. We are not listing all the features because of large size.

For this research work, we have used 2.27 GHz Intel core i3 CPU workstation with 3GB of RAM.

5. METHODOLOGY AND RESULT ANALYSIS

5.1 Methodology

In this research work, Weka toolkit, [16] which is a well-known data mining tool, is used for implementing classification of various internet applications into educational and non-educational classes with three machine learning algorithms. These three machine learning algorithms are RBF, C4.5 decision tree and Bayes Net Classifier. In this implementation dataset of 1740 samples is utilized. In this dataset, 1500 samples are used for training and 240 samples are used for testing purpose.

In this research work, classification accuracy, training time, recall and precision values [2], [10] of individual samples are considered for performance evaluation of these three machine learning 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. Its value ranges from 0 to 1.

5.2 Results and Analysis

Each ML algorithm is trained using training data set and then tested for their performance using test data set. Table 1 shows classification accuracy and training time of RBF, C4.5 and Bayes Net machine learning classifiers. It is clear from this table and figure 2 that maximum classification accuracy is provided by Bayes Net classifier which is 76.67 %. From table 1, it is evident that training time of Bayes Net classifier is 2 second only. From this table, it is obvious that RBF classifier is a slow classifier with training time of 4 seconds and its classification accuracy is also very less i.e. only 65.83%. Therefore, RBF classifier is not suitable for this classification purpose.

Table 1. Classification Accuracy and Training Time of three ML classifiers

Machine Learning Classifiers	RBF	C4.5	Bayes Net
Classification Accuracy (%)	65.83	69.58	76.67
Training Time (Seconds)	4	2	2

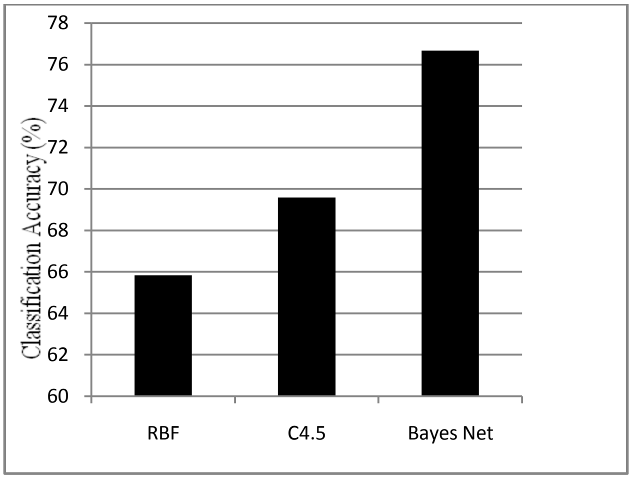

Figure 2: A comparison of classification Accuracy of three ML Classifiers

From these results, it is evident that Bayes Net gives better performance in terms of classification accuracy as compared to RBF and C4.5 classifiers. Figure 3 and 4 show recall and precision values of RBF, C4.5 and Bayes Net classifiers for educational and non-educational internet application categories. Bayes Net gives 96% recall value for educational, and 72% recall value for Non- educational applications. Similarly, it gives 85% precision for educational and 63% Precision value for Non-educational applications. Thus it is again clear that Bayes Net gives better performance in terms of Recall and precision for both educational as well as non-educational applications as compared to other two classifiers.

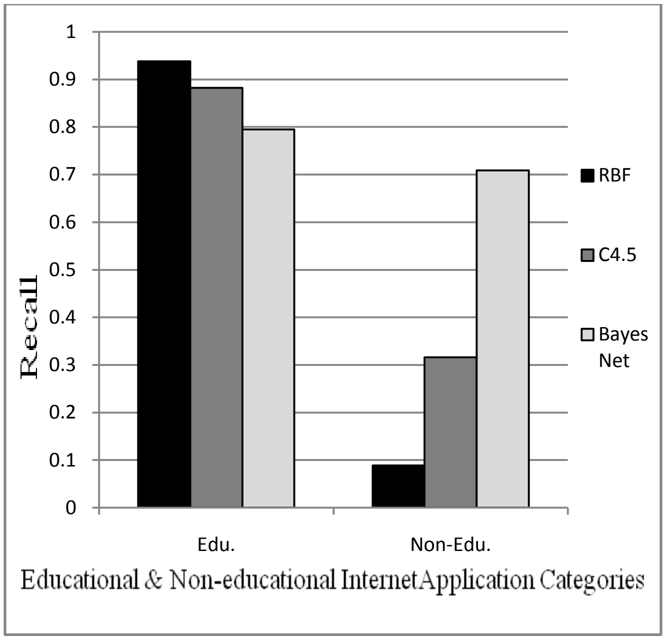

Figure 3: Recall Value of three ML classifier

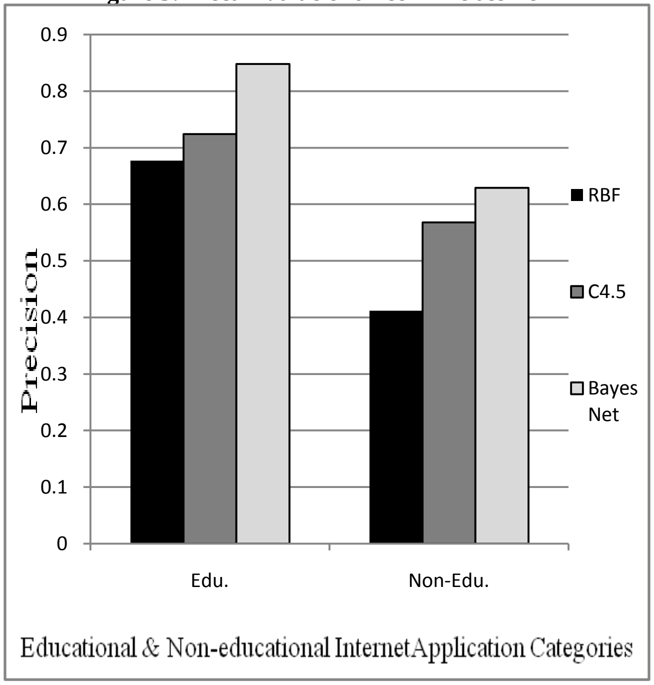

Figure 4: Precision Value of three ML classifiers

From this analysis, it is evident that Bayes Net is a very good classifier for classification of various internet applications into educational and non-educational categories. This classifier gives good performance in terms of classification accuracy, training time, recall and precision of individual samples.

Though, the Bayes Net outperforms the other ML algorithm for this intended classification, but its performance could be expected to improve further by increasing the number of training samples in training data set.

6. CONCLUSIONS AND FUTURE SCOPE

In this paper, firstly internet traffic related to various educational and non-educational internet applications has been captured using Wireshark software which is a packet capturing tool and a dataset has been developed from it. After that, Internet traffic is classified using three machine learning classifiers: RBF, C4.5 and Bayes Net. Results show that Bayes Net gives better classification of internet traffic data in terms of classification accuracy, training time of classifiers, recall and precision values of classifiers for samples. Classification accuracy provided by Bayes Net classifier is 76.67% which is very high as compared to that of other two classifiers. Thus it is evident that Bayes Net is an efficient machine learning technique for classification of internet traffic into educational and non-educational categories.

To improve the performance of ML classifier, our future work will include:

- An increase in number of samples in the training data set.
- An increase in the capture duration for the training data set, so that a significant variation in the feature values for different classes could be observed.
- Extraction of more number of features and selecting most relevant features for intended classification.

Also, various websites related with internet banking, research areas, jobs related websites etc. can also be included under the category of educational websites in future.

In this research work, internet traffic dataset has been developed by considering packet flow duration of 1 minute for each application which is still very large, as far as test data set is concerned. This flow duration can be further reduced in order to make this classification 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.

7. REFERENCES

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

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

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

[4] http:/www.iana.org/assignments/port numbers.

[5] Andrew W. Moore, Denis Zuev, Michael L. Crogan. August 2005. Discriminators for use in flow-based classification. Queen Mary University of London, Department of Computer Science, RR-05-13, ISSN 1470-5559.

[6] S. Zander, T. Nguyen and G. Armitage. November 2005. Automated Traffic Classification and Application Identification using Machine Learning. In LCN'05, Sydney, Australia.

[7] S. Zander, T. Nguyen and G. Armitage. March 2005. Self-Learning IP Traffic Classification Based on Statistical Flow Characteristics. In PAM'05, Boston, USA.

[8] Murat Soysal, Ece Guran Schmidt. 2010. Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison. Performance Evaluation Elsevier Journal, Vol. 67, pp. 451-467.

[9] Indra Bhan Arya and Rachna Mishra. 2011. Internet Traffic Classification: An Enhancement in Performance using Classifiers Combination. International Journal of Computer Science and Information Technologies, Vol. 2 (2), pp. 663-667.

[10] Kuldeep Singh and Sunil Agrawal. 2011. Internet Traffic Classification using RBF Neural Network. In Proceedings of International Conference on Communication and Computing technologies (ICCCT-2011), (Jalandhar, Punjab, India) 39-43.

[11] Yongli Ma, Zongjue Qian, Guochu Shou, Yihong Hu. 2008. Study on Preliminary Performance of Algorithms for Network Traffic Identification. 978-0-7695-3336-0/08 $25.00 © IEEE DOI 10.1109/CSSE.1277, pp.629-633.

[12] Kuldeep Singh and Sunil Agrawal. 2011. Comparative Analysis of five Machine Learning Algorithms for IP Traffic Classification. International Conference on Emerging Trends in Networks and Computing Communications (ENCTT-2011), Udaipur, Rajasthan, India.

[13] Y.L. Chong and K. Sundaraj. 2009. 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).

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

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

[16] Weka website 2011. http://www.cs.waikato.ac.nz/ml/weka/

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

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

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