I.J. Intelligent Systems and Applications, **2012, 8, 37-45**
Published Online July 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2012.08.05

Internet Traffic Classification for Educational Institutions Using Machine Learning

Jaspreet Kaur
University Institute of Engineering & Technology, Panjab University, Chandigarh (India)-160014
Email: jasgill.89@gmail.com

Sunil Agrawal
University Institute of Engineering & Technology, Panjab University, Chandigarh (India)-160014
Email: s.agrawal@hotmail.com

B.S.Sohi
Chandigarh Group of Colleges, Gharuan, Punjab (India)
Email: bssohi@yahoo.com

***Abstract*—** In recent times machine learning algorithms are used for internet traffic classification. The infinite number of websites in the internet world can be classified into different categories in different ways. In educational institutions, these websites can be classified into two categories, educational websites and non-educational websites. Educational websites are used to acquire knowledge, to explore educational topics while the non-educational websites are used for entertainment and to keep in touch with people. In case of blocking these non-educational websites students use proxy websites to unblock them. Therefore, in educational institutes for the optimum use of network resources the use of non-educational and proxy websites should be banned. In this paper, we use five ML classifiers Naïve Bayes, RBF, C4.5, MLP and Bayes Net to classify the educational and non-educational websites. Results show that Bayes Net gives best performance in both full feature and reduced feature data sets for intended classification of internet traffic in terms of classification accuracy, recall and precision values as compared to other classifiers.

***Index Terms*—** Internet traffic classification, Educational websites, Non-educational websites, Proxy websites, Machine Learning, Features

I. Introduction

Over the recent years there has been a drastic growth in internet users both for educational and non-educational purposes. Besides the traditional applications of internet like e-mail, web, the new applications like gaming, P2P have done a major contribution in the rise of the internet traffic. Due to the use of a number of internet applications by users in different fields, the internet traffic increases day by day. Traffic classification is a process which categorizes the internet traffic into various classes based on various parameters e.g. port number or payload or protocol etc. Internet service providers as well as enterprise networks require the ability to accurately identify the different applications, for a range of uses, including security, monitoring to accounting, to detect network intrusions, to detect network misuse by internal and external users and many more. 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 e.g. www.ieeeexpore.org. Similarly non-educational websites can be used for entertainment and to keep in touch with people and to get to know more people e.g. www.bittorrent.com.

In our research work internet traffic is to be classified into two classes, one for educational websites and another for non-educational websites. 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. To optimize the network performance at educational level, this type of classification is of prime importance by which we can handle various network related issues like bandwidth provisioning, resource provisioning, efficient use of network resources, preventing the students from wasting their time in surfing of non-educational websites. So, for the optimum use of network resources the use of non-educational websites should be banned in educational institutions while only the educational websites should be allowed to open.

Moreover the use of non-educational websites can be banned in a number of ways e.g. IP blocking blocks the connection between a website and certain IP addresses or ranges of addresses. IP ban is often used to prevent a disruptive member from accessing a non-educational

website. Web security guard, Routers, Firewalls, Internet filter or URL filter etc. can be used to block the non-educational websites. On the other hand, there are many ways to unblock these websites or to bypass the internet filters e.g. by using web proxies etc. Unblocking a website is a process of gaining access to a particular website which is blocked. In case of IP banning the solution for that is to change your IP address either permanently or temporarily. In case of blocking through Routers or Firewalls, the use of proxy sites will unblock the blocked websites. Proxies allow users to make indirect network connections to other computer network services. There are 3 types of HTTP proxies:

(1) Fully anonymous (elite) proxies: Such proxies do not change request fields and look like real browser. Our real IP is also hidden of course. People that administrate internet servers will think that you are not using any proxies.

(2) Anonymous proxies: They also do not show a real IP but change the request fields, so it is very easy to detect that proxy while log analyzing. Nothing really matters, but some server administrators restrict the proxy requests.

(3) Transparent proxies (not anonymous, simply HTTP): They change the request fields; also they transfer the real IP. Such proxies are not applicable for security and privacy while surfing on the web. We can use them only for network speed improvement [1].

In our research work, we have captured internet traffic from these proxy websites also in order to block these proxy websites from unblocking the non-educational websites. Historically, IP traffic classification techniques were 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. Due to disadvantages of direct packet inspection techniques the research community is now looking for the ML (machine learning) techniques in which, 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.) [4].

In this paper, internet traffic datasets for both educational and non-educational websites have been developed. The dataset for the proxy websites has also been developed and is kept under the category of non-educational websites because these proxy websites are used to access the blocked non-educational websites so the use of these proxy websites should also be banned in educational institutions. From this dataset, a reduced feature data set is also developed using CFS and CON feature reduction algorithms. Then using this full feature and reduced feature datasets, five ML algorithms have been employed 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 [5]. Performance of all these classifiers is analysed on the basis of classification accuracy, training time of classifiers, recall and precision values of classifiers [2].

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.

II. Related Work

There has been much recent work in the field of traffic classification. Various researchers have shown their interest in internet traffic classification over last few years. For this research work, numbers of research papers have been reviewed. Some previous work done in this field by some researchers is discussed as follows:

In [6] 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 [7] 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 [8] Shijun Huang et al. have demonstrated the statistical features based approach to classify internet traffic using supervised ML. The simplified statistical features and the easy-to-use k-Nearest Neighbor (KNN) estimator result in lower space and time complexity, which is worth mentioning. They carried out several data sets including 9 flows of MAIL, 100 flows of

WWW, 34 flows of BULK, 100 flows of IM, 100 flows of P2P and 5 flows of STREAM (full-flow) are collected in the way mentioned in section III, all of which are used to train the k-Nearest Neighbor estimator. They inferred that the classifier model works perfectly when classifying only MAIL, WWW and BULK flows. But with IM flows added, classification results of MAIL flows drop greatly, which breaks the principle of fairness in KNN algorithm. More problems are discovered when P2P flows are added.

In [9] 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 %.

In [10] Agrawal and Sohi demonstrated that P2P applications supposedly constitute a substantial proportion of today's Internet traffic. The ability to accurately identify different P2P applications in internet traffic is important to a broad range of network operations including application-specific traffic engineering, capacity planning, resource provisioning, service differentiation, etc. In this paper, they presented a Neural Network approach that precisely identifies the P2P traffic using Multi-Layer Perceptron (MLP) neural network. This paper has demonstrated the selection of features and successful application of Multi Layer Perceptron (MLP) neural network for P2P traffic identification. Their proposed 'universal' feature set is more effective because it could achieve an improvement of 1.98% in mean precision and 27.81% in mean recall over the feature set selected from traditional method. A very large increase in Recall is noteworthy since high precision is meaningful only when the classifier achieves high value of recall.

III. Machine Learning Concepts

In this research paper, five well-known machine learning algorithms are employed using Weka [11] which is reported in different research papers to be performing well in most of the applications. Also two feature reduction algorithms are employed. These machine learning algorithms and feature reduction algorithms are discussed in brief as follows:

A. Machine Learning Algorithms

Five ML algorithms used are as follows:

(1) Naïve Bayes

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumption. Simply, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable.

A Naïve-Bays' (NB) ML algorithm [12, 13] has a simple structure in which the class node is the parent node of all other nodes. Fig.1 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.

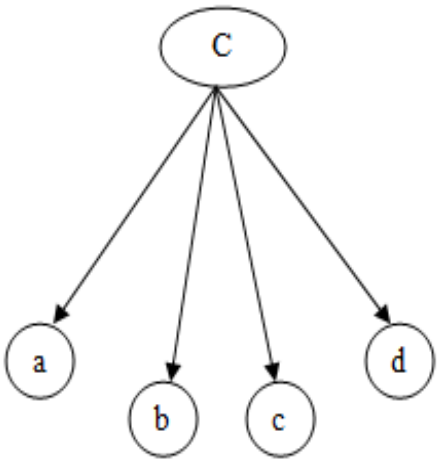

Fig.1 Naïve-Bays classifier

An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.

(2) Radial Basis Function Neural Network

Radial basis function (RBF) [7, 14 and 15] networks typically have three layers: an input layer, a hidden layer with a non-linear RBF activation function and a linear output layer. Radial Basis Function (RBF) Neural Network 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 Fig.2. 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:

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

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

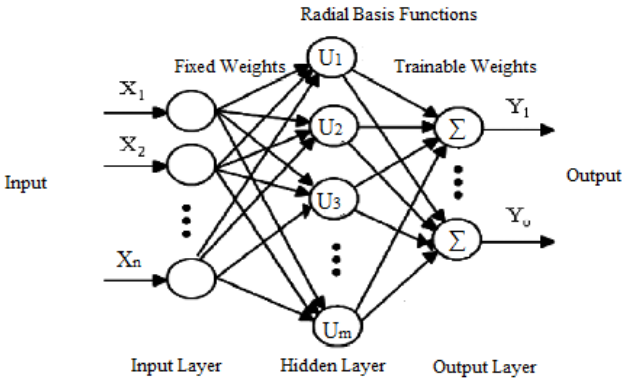

Fig.2 Radial Basis Function Neural Network

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

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

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). They are used in function approximation, time series prediction, and control.

(3) C 4.5 Algorithm

C4.5 is a well-known decision tree Machine Learning algorithm used to generate Univariate decision tree [16]. 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. C4.5 has made a number of improvements to ID3 like it can handle both continuous and discrete attributes, it can handle training data with missing attribute values, it can also handle attributes with differing costs etc. In present research work, C4.5 algorithm has been used for internet 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 [11].

(4) Multilayer Perceptron

A multilayer perceptron (MLP) is a feed forward artificial neural network model which maps a set of input data onto a set of appropriate output. An MLP model consists of multiple layers of nodes with each layer fully connected to the other one. It is also known as Back Propagation 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.3. It consists of input layer having i neurons, hidden layer having j neurons and output layer having k neurons.

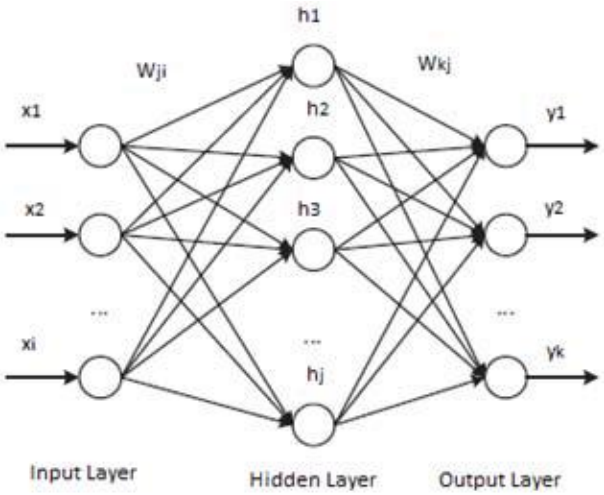

Fig.3 Multilayer Perceptron

MLP's are applicable in many fields. Currently, they are most commonly used in speech recognition, image recognition, and machine translation software. In general, their most important use has been in the growing field of artificial intelligence. 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].

(5) Bayes Net Algorithm

A Bayesian network or Bayes network [5, 12] is popularly called as belief network. It is a probabilistic graphical model which represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). In this model, each node represents random variables; nodes which are not connected represent variables which are conditionally independent of each other 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 [5, 11].

B. Feature Reduction Algorithms

Feature selection, also known as feature reduction, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for building robust learning models. However, training the classifier with maximum number of features obtained is not always the best option, as the irrelevant or redundant features can cause negative impact on a classifier's performance [17] and at the same time, the build classifier can be computationally complex. Two feature reduction methods that have been chosen for this application are CFS [18] and CON [19] as these methods have been widely used by researchers for feature reduction.

IV. Internet Traffic Dataset

In this research work, Wireshark, [20], 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. Data for proxy websites has also been captured and is kept under the category of non-educational websites because the proxy websites are used to unblock the non-educational websites. Therefore, to avoid this unblockage and for the optimum use of network resources in the educational institutions these proxy websites should also be blocked. 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. In this process of developing datasets, two datasets are obtained: one is full feature dataset and another is reduced feature dataset [4]. In full feature data set, 108 features are extracted for each website using MATLAB [21] out of which, six features are extracted directly from statistics summary of Wire shark. While, other 102 features are extracted for TCP and UDP conversations of Wire shark.

In this process of packet capturing and feature extraction, a dataset of 497 samples is developed by performing feature extraction of traffic traces using MATLAB code. Then this dataset is divided into training and testing data sets. In training dataset 350 samples are taken while the testing dataset contains 147 samples. The 350 training samples are then up sampled to 3500 samples using Weka tool. Each sample is characterized by 108 features which mainly consists 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. In case of reduced feature data set, two data sets are developed from CFS and CON algorithms. In CFS algorithm, cfsSubsetEval evaluator and Best First search in attribute selection filter of Weka tool [11] is used, while in CON algorithm, consistency Subset Eval evaluator and Best First search in attribute selection filter of Weka tool [11] is used. For this research work, we have used 2.27 GHz Intel core i3 CPU workstation with 3GB of RAM.

V. Experimental Implementation and Result Analysis

In this section, method used and result analysis are discussed.

A. Methodology

In this research work, Weka toolkit, [11] which is a well known data mining tool, is used for implementing classification of various internet applications into educational and non-educational classes with five machine learning algorithms. These five machine learning algorithms are Naïve Bayes, RBF, C4.5 decision tree, MLP and Bayes Net Classifier. In this implementation, dataset of 3647 samples is utilized. In this dataset, 3500 samples are used for training and 147 samples are used for testing purpose. In this research work, classification accuracy, training time, recall and precision values [2, 7] of individual samples are considered for performance evaluation of these five machine learning classifiers for full feature as well as reduced feature data sets. 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.

B. Result 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, training time, Recall and Precision values of Naïve Bayes, RBF, C4.5, MLP and Bayes Net machine learning classifiers. It is clear from this Table that maximum classification accuracy is provided by Bayes Net classifier which is 96.6 %. From Table 1, it is evident that training time of Bayes Net classifier is 17 seconds. Also, from this Table, it is obvious that MLP and RBF classifier are slow classifiers with training time of 264 seconds and 38 seconds respectively. Even MLP has classification accuracy of 91.84% but it has high training time of 264 seconds also. The training time is minimum for Naïve Bayes classifier i.e. 8 seconds only but it has minimum classification accuracy of only 72.79%. Therefore, MLP and Naïve Bayes classifiers are not suitable for this classification purpose. From these results, it is evident that Bayes Net gives better performance in terms of classification accuracy as compared to other classifiers. Bayes Net gives 96.6% recall value. Similarly, it gives 96.7% precision values. Thus it is again clear that Bayes Net gives better performance in terms of Recall and precision values as compared to other classifiers.

Table 1 Classification Accuracy, Training Time, Recall and Precision values of Five ML Classifiers

Machine Learning Classifiers	Naïve Bayes	RBF	C 4.5	MLP	Bayes Net
Classification Accuracy (%)	72.79	75.51	89.11	91.84	96.6
Training Time (Seconds)	8	38	24	264	17
Recall	0.728	0.755	0.891	0.918	0.966
Precision	0.869	0.876	0.888	0.934	0.967

Further, to reduce the feature set, we apply CFS and CON feature reduction techniques using the Best first search method, as it is a commonly used method and yields good results. CFS feature reduction algorithm results in 11 features and CON feature reduction algorithm results only in 3 features. Table 2 shows the list of 11 features obtained using CFS. With this feature set, performance of the chosen five ML algorithms is analyzed with the selected 11 features.

Table 2 List of CFS Features

Feature Number	Feature Name
1	Min. Duration of conversation (TCP)
2	No. of conversation (UDP)
3	Max. of no. of packets in conversation (UDP)
4	Mean of bytes in conversation (UDP)
5	Max. no. of packets from A to B in conversation (UDP)
6	Total no. of packets from A to B in conversation (UDP)
7	Min. of relative time between start of capturing and start of conversation (UDP)
8	Max. duration of conversation (UDP)
9	Total duration of conversation (UDP)
10	Min. Bit rate (bps) from A to B of conversation. (UDP)
11	Internet Application Category

Table 3 shows classification accuracy, training time, Recall and Precision values of Naïve Bayes, RBF, C4.5, MLP and Bayes Net machine learning classifiers for CFS feature reduction algorithm. It is clear from this Table that maximum classification accuracy is provided by Bayes Net classifier which is 97.96 % and it is clear that classification accuracy of Bayes Net improves with reduced feature data set using CFS as compared to full feature data set. Also the training time for Bayes Net reduces from 17 to 3 seconds only. Therefore it is clear that Bayes Net performance is improved by reducing the features using CFS algorithm. Also performance of RBF and MLP has improved with CFS algorithm. But on the other hand, the classification accuracy of C 4.5 and Naïve Bayes has been reduced with a large reduction in training time.

Table 3 Classification Accuracy, Training Time, Recall and Precision values of Five ML Classifiers for CFS Algorithm

Machine Learning Classifiers	Naïve Bayes	RBF	C 4.5	MLP	Bayes Net
Classification Accuracy (%)	50.34	91.16	85.71	93.88	97.96
Training Time (Seconds)	2	9	4	86	3
Recall	0.503	0.912	0.857	0.939	0.98
Precision	0.813	0.93	0.858	0.947	0.98

Now the original feature set is subjected to CON feature reduction method and 3 features are obtained as mentioned in Table 4.

Table 4 List of CON Features

Feature Number	Feature Name
1	Time between first and last packet
2	Total duration of conversation (UDP)
3	Internet Application Category

Table 5 shows classification accuracy, training time, Recall and Precision values for five machine learning classifiers for CON feature reduction algorithm. It is clear from Table 5 that performance of CON is not better than CFS as well as full feature data set

performance in terms of classification accuracy. Only the training time has been reduced as compared to both of them.

Table 5 Classification Accuracy, Training Time, Recall and Precision values of Five ML Classifiers for CON Algorithm

Machine Learning Classifiers	Naïve Bayes	RBF	C 4.5	MLP	Bayes Net
Classification Accuracy (%)	81.63	85.71	79.59	74.83	91.16
Training Time (Seconds)	1	3	2	28	1
Recall	0.816	0.857	0.786	0.748	0.912
Precision	0.886	0.884	0.786	0.56	0.913

Fig.4 shows the classification accuracy for full feature dataset, reduced feature dataset for both CFS and CON algorithms.

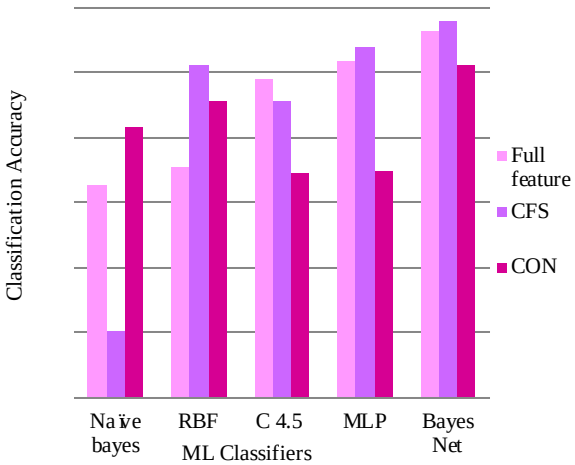

Fig.4 A comparison of classification Accuracy of five ML Classifiers

Fig.5 shows recall value for full feature dataset, reduced feature dataset for both CFS and CON algorithms.

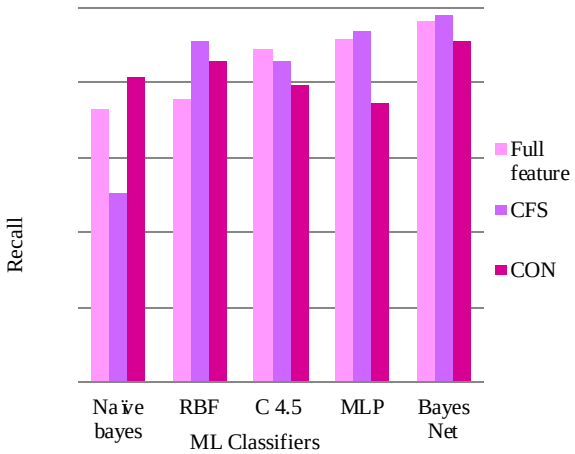

Fig.5 A comparison of recall value of five ML Classifiers

Fig.6 shows precision value for full feature dataset, reduced feature dataset for both CFS and CON algorithms.

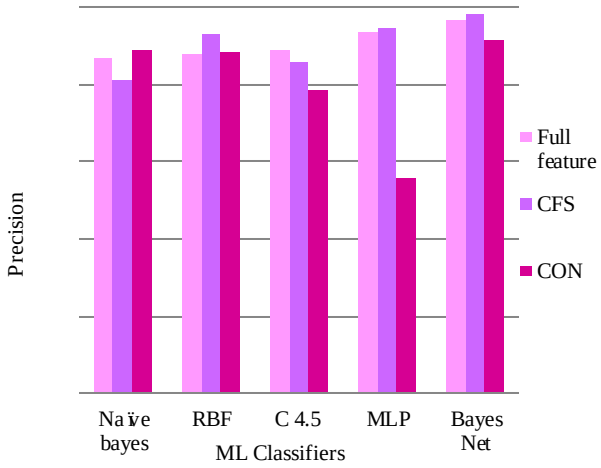

Fig.6 A comparison of precision value of five ML Classifiers

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, 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. Also 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.

VI. Conclusion

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. Data for proxy websites has also been captured and is kept under the category of non-educational websites because the proxy websites are used to unblock the non-educational websites. Therefore, to avoid this unblockage and for the optimum use of network resources in the educational institutions these proxy websites should also be blocked. After that, Internet traffic is classified using five machine learning classifiers: Naïve Bayes, RBF, C4.5, MLP and Bayes Net. Results show that Bayes Net gives better classification of internet traffic data in terms of classification accuracy, recall and precision values of classifiers for samples. Classification accuracy provided by Bayes Net classifier is 96.6% which is very high as compared to that of other 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.
- Decreasing the capture duration to make the data set more real time compatible.
- Extraction of more number of features and selecting most relevant features for intended classification.
- Training time can be reduced to reduce the computational complexity.

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.

References

[1] en.wikipedia.org/wiki/proxy_list

[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] Andrew W. Moore, Denis Zuev and 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.

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

[6] Murat Soysal and 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.

[7] 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.

[8] Shijun Huang Kai Chen Chao Liu, Alei Liang and Haibing Guan. (2009). A Statistical-Feature-Based Approach to Internet Traffic Classification Using Machine Learning. ©2009 IEEE 9781-4244-3941-6/09/$25.00

[9] 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.

[10] S. Agrawal and B. S. Sohi. (2011). Generalization and Optimization of Feature Set for Accurate Identification of P2P Traffic in the Internet using Neural Network. WSEAS TRANSACTIONS on COMMUNICATIONS.

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

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

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

[14] 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).

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

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

[17] N. Williams, S. Zander and G. Armitage. (2006). A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification. ACM SIGCOMM Computer Communication Review, vol. 36, pp. 7-15.

[18] M. Dash and H.Liu. (2003). Consistency-based Search in Feature Selection. Artificial intelligence, vol. 151, pp. 155-176.

[19] M. Hall. (1998). Correlation-based Feature Selection for Machine Learning. PHD Thesis,

Deptt of Computer Science, Waikato University, Hamilton, New Zealand.

[20] Wireshark. Available: http:// www.wireshark.org/

[21] MATLAB. Available:www.mathworks.com

Sunil Agrawal received his B.E. degree in Electronics & Communication in 1990 from Jodhpur University in Rajasthan, India and M.E. degree in Electronics & Communication in 2001 from Thapar University in Patiala, India.

He is Assistant Professor at the University Institute of Engineering & Technology in Panjab University, Chandigarh, India. He has 20 years of teaching experience (undergraduate and postgraduate classes of engineering) and has supervised several research works at masters level. He has 25 research papers to his credit in national and international conferences and journals.

The author's main interests include applications of artificial intelligence, QoS issues in Mobile IP, and mobile ad hoc networks.

Jaspreet Kaur is currently doing M.E. from University Institute of Engg. and Technology, Panjab Universisty, Chandigarh. She has received her B.Tech degree from Punjab Technical University, Jalandhar, India.

Her interests include Neural Networks and Embedded System.

Balwinder S. Sohi is currently working as Campus Director at CGC – a group of colleges in Engineering & Technology. He has a long administrative experience as Director, University Institute of Engineering & Technology – a premier engineering institute of Panjab University at Chandigarh, India. He has served at various faculty positions as Professor, Assistant Professor and Lecturer, during his long professional carrier of 38½ years (including 4 years in research organizations). Graduated from Panjab Engineering College, Chandigarh, he has attained his masters and PhD degrees from Panjab University, Chandigarh.

He has guided the research works at masters and PhD levels and has more than 70 national and international research papers to his credit. He has been responsible in setting up various facilities in the field of Electronics & Communication, through sponsored projects from agencies like MHRD, AICTE, DIT etc. He has been Dean of Engineering & Technology at Panjab University, Chandigarh. He has contributed to technical education in various capacities at different fora like AICTE etc. He has been honored twice by Institute of Engineers, Kolkata, India, for best research work.

How to cite this paper: Jaspreet Kaur,Sunil Agrawal,B.S.Sohi,"Internet Traffic Classification for Educational Institutions Using Machine Learning", International Journal of Intelligent Systems and Applications(IJISA), vol.4, no.8, pp.37-45, 2012. DOI: 10.5815/ijisa.2012.08.05

