I.J. Intelligent Systems and Applications, **2013, 03, 83-93**
Published Online February 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2013.03.09

A Near Real-time IP Traffic Classification Using Machine Learning

Kuldeep Singh
RIMT Maharaja Aggrasen Engineering College, Mandi Gobindgarh, Punjab (India)
kuldeepsinghbrar87@gmail.com

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

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

Abstract— With drastic increase in internet traffic over last few years due to increase in number of internet users, IP traffic classification has gained significant importance for research community as well as various internet service providers for optimization of their network performance and for governmental intelligence organizations. Today, traditional IP traffic classification techniques such as port number and payload based direct packet inspection techniques are rarely used because of use of dynamic port number instead of well-known port number in packet headers and various cryptographic techniques which inhibit inspection of packet payload. Current trends are use of machine learning (ML) techniques for IP traffic classification. In this research paper, a real time internet traffic dataset has been developed using packet capturing tool for 2 second packet capturing duration and other datasets have been developed by reducing number of features of 2 second duration dataset using Correlation and Consistency based Feature Selection (FS) Algorithms. Then, five ML algorithms MLP, RBF, C4.5, Bayes Net and Naïve Bayes are employed for IP traffic classification with these datasets. This experimental analysis shows that Bayes Net is an effective ML technique for near real time and online IP traffic classification with reduction in packet capture duration and reduction in number of features characterizing each application sample with Correlation based FS Algorithm.

Index Terms— IP Traffic Classification, Machine Learning Techniques, Feature Selection, Packet Capture Duration, Classification Accuracy, Training Time

I. Introduction

With rapid growth in internet users in major sections of society over few years, internet traffic is going to be increased at drastic rate. This increase in internet traffic is due to the use of variety of internet applications by users in their day to day life such as www, e-mail, peer to peer (P2P) applications, FTP applications, instant messaging, audio/video calls, multimedia applications etc. Therefore, IP traffic classification is a necessary task for various internet service providers (ISPs) in order to optimize network performance by solving various network monitoring and management problems such as available bandwidth planning and provisioning, measure of QoS of various internet applications, billing information for the subscribers on a particular network, and any severe problem which degrades the performance of network etc [1, 2]. Now a day, it is also being utilized by various governmental intelligence agencies in order to solve various security related issues.

The significant contribution in internet traffic is done by peer to peer (P2P) applications which mainly include Bit Torrent, Emule, Kaaza etc. These applications lead to 80% rise in internet traffic [2]. Now a day, various multimedia websites such as Youtube also contribute a major portion in internet traffic. 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 chatting which is a form of instant messaging. These applications are again a major reason to rise in internet traffic.

Traditionally, various IP traffic classification techniques have been based upon direct inspection of packets flowing through the network [1]. These techniques are payload based and port number based packet inspection techniques. In payload based technique, payload of few internet packets are analyzed in order to identify any particular internet application which is not possible now a days because of use of cryptographic techniques used for encryption of 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 internet packets which are reserved by IANA (Internet Assigned Numbers Authority) for particular applications e.g. port number 80 is reserved for web based applications [3]. Unfortunately, this method is also rarely used due to the use of dynamic port numbers instead of well-known port numbers for various applications.

Current trends are use of Machine Learning (ML) techniques [1,18] for IP traffic classification in which ML models or networks are trained using a set of previous examples consisting of training input and target pair and then those trained networks are used for predicting classes of unknown test samples. In this research article, a real time internet traffic dataset has been developed. In this dataset, packet capturing duration is taken as 2 seconds. With this dataset, five well-known 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 Na ïve Bayes Algorithm [18]. Performance of all these classifiers has been evaluated on the basis of classification accuracy, training time of classifiers, recall and precision values of classifiers for individual internet applications and number of features used to characterize internet application samples [1,13]. In order to make IP traffic classification more real time compatible, reduced feature datasets have been developed from full feature dataset using Correlation based [12] and Consistency based [13] Feature Selection Algorithms. Using these datasets, IP traffic classification has been implemented again with same five ML algorithms. Results reveal that Bayes Net gives very high accuracy with smaller training time making it more suitable for near real time IP traffic classification with reduction in number of features characterizing each internet application using Correlation based feature selection algorithm.

The rest of the article is organised as follows: section II gives some information about background and related work in IP traffic classification field. Section III includes some introductory information about ML classifiers and Feature Selection algorithms used in this research work. Section IV gives overview of datasets and internet traffic classes. Methodology and result analysis is given in section V. Section VI provides conclusions and future scope for other researchers who are willing to work in this field.

II. Background and Related Work

A lot of research work has been done in the area of IP traffic classification by considering various internet application types and numbers of classification techniques have been suggested by researchers in this field. However the majority of them are based on port number, payload based and protocol behavior based. In current scenario machine learning techniques are widely used for IP traffic classification. All these approaches are explored in detail in following subsections.

2.1 Port Number Based IP traffic classification

This method is used to classify internet traffic with the help of well-known port number in TCP/UDP packet headers which is reserved for particular internet application by Internet Assigned Numbers Authority (IANA) [3]. Initially, this method was very effective and easy to implement for real time IP traffic classification. However, nowadays, various internet applications such as P2P do not use well-known ports to avoid being detected or applications such as FTP in passive mode, change their ports dynamically [4]. Other applications such as multimedia streaming and online internet games also use dynamic port numbers instead of well-known port numbers. Some standard internet services or applications can potentially run on nonstandard or dynamic port numbers not reserved by IANA to circumvent policy or operating system access control restrictions. Thus, port-based IP traffic classification cannot produce true results. The classification accuracy for port-based approaches is reported to be between 50% and 70% [5]. Thus in today’s scenario, this technique becomes ineffective for IP traffic classification.

2.2 Payload Based IP traffic classification

This method is used to inspect internet traffic packet payloads to search for exact signatures of known applications. Previous studies show that payload based signature identification method work very well with high classification accuracy for the current internet traffic including many of P2P traffic.

In [6], Subhabrata Sen et al. have provided an efficient approach for identifying the P2P application traffic through application level signatures. They have examined the performance of this application-level identification approach using five popular P2P protocols namely Gnutella, eDonkey, Direct Connect, Bit Torrent and Kazaa protocols. The measurements in this article show that this technique achieves less than 5 % false positive and false negative ratios in most cases.

But this classification technique is still not widely acceptable because of its many limitations. First of all, this method can only identify internet traffic for which signatures are available and are unable to classify any other traffic [4]. Secondly, packet payload analysis requires very high storage capacity and computational power because it has to analyze full packet payload. One important issue behind ineffectiveness of this technique is privacy policies of government which do not allow any unaffiliated third party to inspect packet payload [1]. Finally, nowadays, various cryptographic techniques are used to encrypt packet payload which leads to failure of this technique for IP traffic classification.

2.3 Host Behaviour Based IP traffic classification

This method is a novel approach to classify internet traffic in various application categories based on observing and identifying patterns of host behavior at the transport layer. The main advantage of this approach is that there is no need for packet payload access and no knowledge of port numbers, hence overcoming the limitations of payload based and port number based IP traffic classification technique.

In [7], Thomas Karagiannis et al. have proposed a novel approach for IP traffic classification, known as BLINd Classification or BLINC in short. In this approach, inherent behaviours of a host are captured at three different levels namely social level, functional level and application level. Social level examines the popularity of the host. The functional level gives information about whether the intended host provides or consumes any particular service. Finally, the application level which is intended to identify the application of the origin is considered. This article shows that this approach can classify approximately 80% -90% of the total number of flows in each trace with 95% accuracy. However this approach cannot be applied for real time classification because of the problem of encrypted header which leads to failure of accessing fields of packet headers and this approach have very slow classification speed.

2.4 Machine Learning Based IP traffic classification

In this method, various machine learning (ML) algorithms are used for IP traffic classification. It uses the concept of statistical analysis based classification. It uses various statistical features related to packet flow such as packet size, number of packets, inter packet arrival time, duration etc for classification purpose [14]. The main advantage of this technique is that there is no need of inspection of packet payload or packet port number. Machine Learning techniques mainly involve Neural Networks, Decision Trees and Bayesian Networks for IP traffic classification.

In [8], Runyuan Sun et al. have designed host based traffic collection platform to order to collect internet traffic of web, P2P and other applications. They have employed three methods for IP traffic classification such as Probabilistic neural network (PNN), RBF neural network and Support Vector Machine (SVM). This article concludes that PNN gives better performance as compared to other two networks. But this research work is limited to web and P2P applications only because other internet applications are not taken into account.

In [9], Li Jun et al. have used machine learning technique for IP traffic classification. In their article, Generic Algorithm has been used for feature reduction in order to reduce training time and computational complexity. They used popular ML algorithms such as TAN, C4.5, Naïve Bayes, Random Forest and distance weighted KNN for traffic classification. This article concludes that C4.5 and Random Forest give remarkable performance for IP traffic classification. But scope of this article is limited to few application samples and their categories. Better performance can be obtained using other ML techniques.

In [10], 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], Singh and Agrawal have developed a real time internet traffic dataset for 2 minute packet capturing duration by considering start and end of each particular application and using attribute selection algorithm, a reduced feature dataset has also been developed. Then five ML algorithms such as MLP, RBF, C4.5, Bayes Net and Naïve Bayes are being utilized for IP traffic classification. Results show that C4.5 algorithm is very effective ML technique for IP traffic classification with classification accuracy in the range of 94% with reduction um number of features characterizing each application sample. But in this work, packet capturing duration is very large. Therefore, this analysis is not very real time compatible. In order to make this analysis more real time compatible, packet capture duration should be as less as possible. In present paper, authors have reduced packet capture duration from 2 minute to 2 second only and then performed IP traffic classification using ML algorithms. Experimental results reveal that reduction in packet capture duration has greater effect in improving classification accuracy and recall and precision values of individual internet applications.

III. ML and FS Techniques

In this research article, five popular machine learning (ML) algorithms are employed for IP traffic classification and two feature selection (FS) algorithms are employed to reduce the numbers of features characterizing each internet application. These algorithms are reported in different research articles to be performing well in most of the applications. These

ML algorithms and FS algorithms are explored in brief as follows:

3.1 Multilayer Perceptron

Multilayer Perceptron (MLP), [15], [16], [22] popularly known as Back Propagation Neural Network, is a multilayer feed forward artificial neural network. 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 weights of the network.

A single hidden layer MLP is shown in figure 1. It consists of input layer which is composed of number of neurons equal to number of features used to characterize a particular input sample, hidden layer which composed of variable or user defined number of neurons and output layer which have number of neurons equal to dimensions of output of the network.

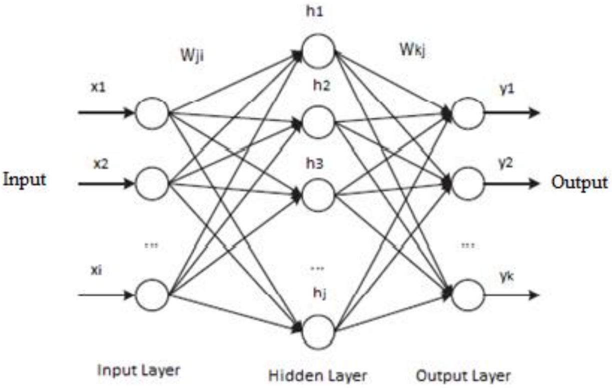

Fig. 1: Multilayer Perceptron

In this research article, single hidden layer MLP is being used for IP traffic classification with learning rate of 0.3 and momentum term of 0.2 [19].

3.2 Radial Basis Function Neural Network

Radial Basis Function (RBF) Neural Network [10, 15, 22] is a multilayer feed forward artificial neural network in which radial basis functions are used 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 structure of RBF neural network is shown in fig 2. In this network, weights for input-hidden layer interconnections are fixed. While, weights for hidden-output layer interconnections are trainable. Each neuron in hidden layer has basis function $U_m(.)$.For any input vector X consisting of X_1, X_2, ……X_n features, the output of this network is given by following input - output mapping function as:

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

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

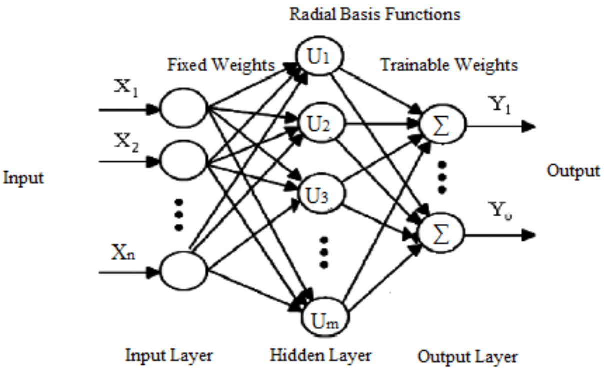

Fig. 2: Radial Basis Function Neural Network

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

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

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).

In this research article, single hidden layer RBF algorithm has been used for IP traffic classification with number of center points in hidden layer equal to 5 2 and again 2 for Datasets of 2 minute packet capture duration , 2 second packet capture duration and reduced feature dataset having 2 second packet capture duration respectively [19].

3.3 C4.5 Algorithm

C4.5 is a popular decision tree ML algorithm which is used to generate Univariate decision tree [17]. It is an extension of Iterative Dichotomiser 3 (ID3) algorithm which is used to find simple decision trees. C4.5 is also known as Statistical Classifier because of its classification capability.

C4.5 makes decision trees from a set of training data samples in similar manner as ID3, using the concept of information entropy. The training dataset consists of large number of training samples characterized by various features and it also consists of target class. C4.5 selects one particular feature of the data sample 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 to make the decision. After that, the C4.5 algorithm repeats its action on the smaller subsets until each subset consists of samples having same target class.

In this research article, 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 both datasets[19].

3.4 Bayes Net

Bayes Net (Bayesian Network), [18, 20] which is popularly known as Belief Network, is a probabilistic graphical model which is used to represent a set of random variables and their conditional dependencies with the help of directed acyclic graph (DAG). 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 research article, Bayes Net algorithm with simple estimator and K2 search algorithm has been used for IP traffic classification [18, 19].

3.5 Naïve Bayes

A Naïve-Bayes ML algorithm [18, 20, 21] is a simple structure which consists of a class node as the parent node of all other nodes. The basic structure of Naïve Bayes Classifier is shown in figure 3. In this figure, C represents main class and a, b, c, and d represents other feature nodes of a particular sample. Other connections are not allowed in a Naïve-Bayes structure. It is easy to construct Naïve Bayes classifier as compared to other classifiers because its structure is provided a priori and therefore it does not require any structure learning procedure. Therefore, this technique uses very small modeling time or training time to model this algorithm for classification purpose. It assumes all the features independent of each other.

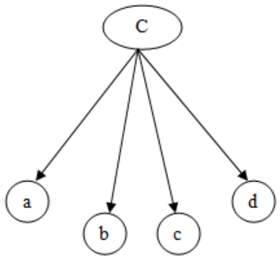

Fig. 3: Naïve Bayes Model

This algorithm performs very effectively over a large number of datasets, especially where the features used to characterize each sample are not properly correlated with each other.

3.6 Correlation based Feature Selection Algorithm

Correlation based Feature Selection (FS) Algorithm, [12] is a popular algorithm used to identify and eliminate those features which are irrelevant for describing particular internet application and redundant in nature. Correlation based FS Algorithm evaluates importance of a subset of features by considering the individual predictive capability of each feature along with the degree of redundancy between them. Subsets of features which are highly correlated with output category, while having low inter-correlation among others are preferred.

In Correlation based FS Algorithm, high scores are assigned to subsets containing attributes that are highly correlated with output category and have low inter-correlation among others. Concept of conditional entropy is considered in order to provide a measure of the correlation between features and class and between features. If H(X) is the entropy of a feature X and H(X|Y) the entropy of a feature X given the occurrence of feature Y the correlation between two features X and Y can then be calculated using the symmetrical uncertainty:

$$C\ (X|Y) = \frac{H(X) - H(X|Y)}{H(Y)} \quad (3)$$

In this algorithm, target class of a data sample is also considered to be a feature. In this research work, Best First search method is used for subset search in the forward and backward directions [12].

3.7 Consistency based Feature Selection Algorithm

Consistency based FS Algorithm, [13] evaluates subsets of features simultaneously and selects the optimal subset. This optimal subset is considered to be the smallest subset of features that can identify samples of a class as consistently as the complete feature set. For determining the consistency of a subset, the combinations of feature values representing a class are given a pattern label. All samples of a given pattern should represent the same class. If two samples of the same pattern represent different classes, then that pattern is taken as inconsistent. This algorithm gives subset of features with very small number of features characterizing each internet application class.

In this research work, Best First search method is used for subset search in the forward and backward directions.

IV. Dataset and Internet Traffic Classes

In this research work, Wireshark, [23] which is a popular packet capturing tool, is used to capture real time internet traffic. Wireshark is a network packet analyzer which is used to capture network packets and to display that packet data as detailed as possible.

A real time internet traffic dataset has been developed in this work. For this dataset, internet traffic packets are captured for the duration of 2 seconds only just by considering on-going middle session of each application. In this packet capturing process, starting and end of applications are not taken into account. This dataset is named as Dataset 1. Since packet capture duration is only 2 second, therefore, numbers of packets captured are very small. But these packets are highly correlated with particular internet application and free of any type of internet noise. Therefore it gives high degree of prediction for IP traffic classification.

The main problem of high training time or model building time by using Dataset 1 is solved by reducing number of features characterizing each internet application. In order to develop reduced feature datasets, Correlation based Feature Selection Algorithm [12] and Consistency based Feature Selection [13] Algorithm of Weka tool have been used . These reduced feature datasets is named as Dataset $2_{\text{Correlation}}$ and Dataset $2_{\text{Consistency}}$ respectively.

In all these datasets, eight internet applications are mainly taken into account such as WWW, e-mail, web media, P2P, FTP data, instant messaging (IM) ,VoIP and Software Updates (SU). WWW type of data samples are obtained from various websites by using Internet Explorer, Google Chrome and Mozilla Firefox internet browsers. For e-mail type data samples, Yahoo Mail and Gmail are taken into account. Main source of Web media type of data samples is Youtube. P2P type of data samples are obtained by using peer to peer applications such as Emulle, μ-torrent, Ares and Shareaza. FTP data samples are obtained by capturing packets from various software and audio/video songs downloading websites. For Instant Messaging type of applications, Yahoo Messenger and Google Talk are considered. VoIP traffic which mainly consists of audio and video conversations on internet is obtained from Yahoo Messenger type of applications. For Software Updates, packets are captured during updating of various Windows software and other software.

Dataset 1, Dataset $2_{\text{Correlation}}$, and Dataset $2_{\text{Consistency}}$ include 2160 data samples in each. In Dataset 1, each application 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 protocol conversations. All these features are extracted from captured packet flows using MATLAB. In Dataset $2_{\text{Correlation}}$, features are reduced to 45 and in Dataset $2_{\text{Consistency}}$, only 8 features are present.

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. Methodology and Result Analysis

5.1 Methodology

In this research work, a general research methodology has been adopted which is shown in figure 4.

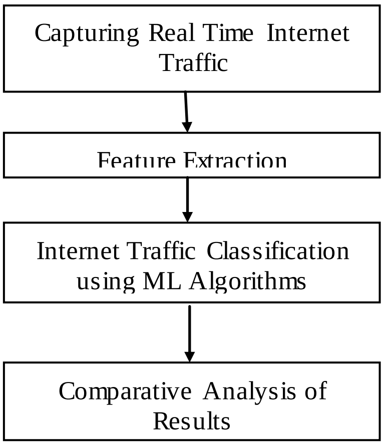
Capturing Real Time Internet Traffic
Feature Extraction
Internet Traffic Classification using ML Algorithms
Comparative Analysis of Results

Fig. 4: Research Methodology

In this work, Weka tool, [19] which is very popular data mining tool, is used for implementing IP traffic classification with five different ML algorithms namely MLP, RBF, C4.5, Naïve Bayes and Bayes Net and two feature selection (FS) algorithms namely Correlation based and Consistency based FS Algorithms. Three different internet traffic datasets namely, Dataset 1 and Dataset $2_{\text{Correlation}}$ and Dataset $2_{\text{Consistency}}$, consisting of 2160 data samples in each dataset, are used in this work. Dataset 1 which is having 2 second packet capture duration and Dataset $2_{\text{Correlation}}$ and Dataset $2_{\text{Consistency}}$ which are reduced feature datasets obtained from Dataset 1are divided into two sets consisting of 2000 data samples for training and 160 data samples for testing purpose in both cases.

In this work, classification accuracy, training time, recall and precision values [1, 11], of individual internet application samples, number of features used to characterize each application sample and packet capture duration are taken into account in order to evaluate performance of these five ML algorithms/classifiers. Some of 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 ML 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 X among all those which were classified as class X. In paper its value ranges from 0 to 1.

5.2 Results and Analysis

Table I shows classification accuracy and training time of five ML classifiers namely MLP, RBF, C4.5, Bayes Net and Naïve Bayes for Dataset 1 which has been developed by considering packet capture duration of 2 seconds only. It is clear from this table and figure 5 that maximum classification accuracy is provided by Bayes Net classifier for Dataset 1 which is 88.125 % with training time or model building time of 0.7 seconds only.

From table I, it is also clear that MLP algorithm gives very poor performance in terms of classification accuracy and training time. Furthermore, classification accuracy is of RBF Neural Network Classifier is also lesser than that of other ML classifiers and its training time is very large as compared to Bayes Net, C4.5 and Naïve Bayes which make it inappropriate for efficient IP traffic classification. Therefore MLP and RBF algorithms are not taken into consideration for further discussion.

Table 1: Classification Accuracy and Training Time of five ML Classifiers for Dataset 1

ML Classifiers	MLP	RBF	C4.5	Bayes Net	Naïve Bayes
Classification Accuracy (%)	27.75	81.25	83.75	88.125	81.875
Training Time (Seconds)	17.79	6.14	1.34	0.7	0.16

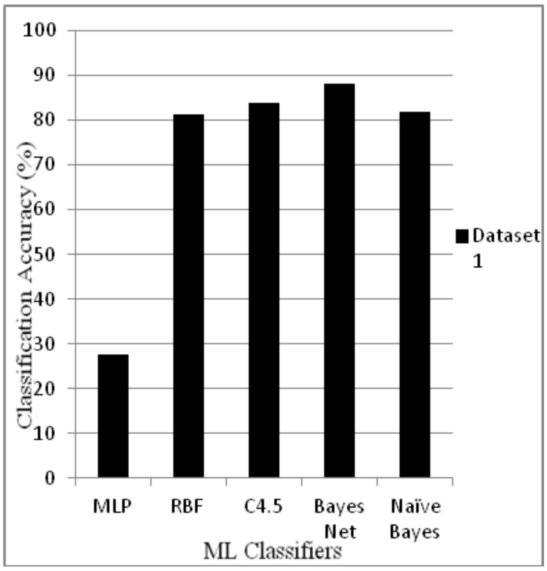

Fig. 5: Classification Accuracy of five ML Classifiers for Dataset 1

From these results, it is clear that Bayes Net gives better performance in terms of classification accuracy and training time as compared to other ML classifiers for Dataset 1which is a full feature dataset. Figure 6 and 7 show recall and precision values of three most accurate ML classifiers i.e. C4.5, Bayes Net and Naïve Bayes for individual internet applications. Bayes gives 100% recall and precision values for most of the applications. Thus it is again clear that Bayes Net give better performance in terms of recall and precision for most of internet applications in case of 2 second duration dataset.

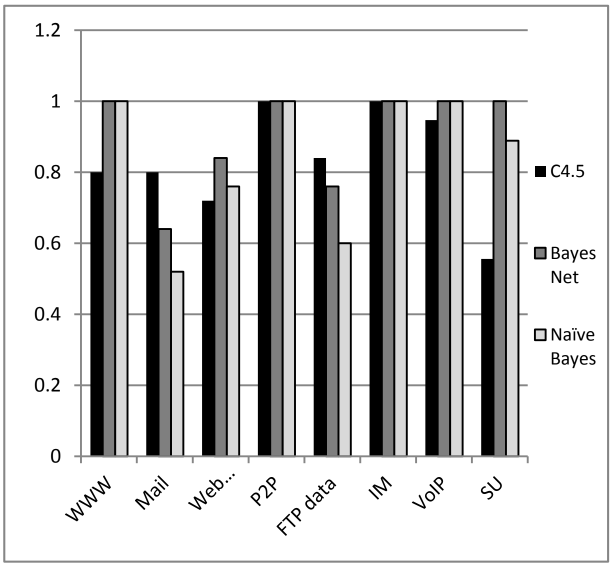

Fig. 6: Comparison of Recall for Dataset 1

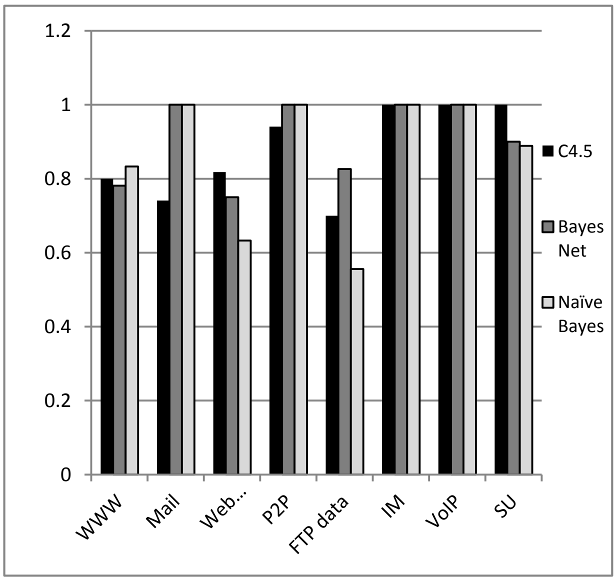

Fig. 7: Comparison of Precision for Dataset 1

Although with dataset 1 which is taken at packet capture duration of 2 seconds, Bayes Net gives better classification performance. But training time or model building time is still very large for these classifiers which make this technique unsuitable for real time and online IP traffic classification. Therefore, it is necessary to reduce the training time of ML algorithms so that real time IP traffic classification can be implemented effectively. This is only possible by reducing the numbers of features characterizing each internet application. In this research work, Correlation based Feature Selection Algorithm and Consistency based Feature Selection Algorithms of Weka tool are employed to reduce number of features in Dataset 1. With Correlation based FS Algorithm, number of features are reduced to 45 features only and this dataset is named as Dataset $2_{\text{Correlation}}$. With In this dataset, all the application samples are characterized by 46 features only. With Consistency based FS Algorithm, number of features are reduced to 8 features only and this dataset is named as Dataset $2_{\text{Consisteny}}$.

Table II and table III show the performance of five ML classifiers in terms of classification accuracy and training time. It is clear from these tables and figure 5 that classification accuracy of all ML algorithms is maximum and training time is reduced to much extent in case of both reduced feature dataset. On the basis of comparative analysis of two FS algorithms, it is revealed from table II and III that Consistency based FS algorithm have only 8 features characterizing each application sample which make it inefficient for effective IP traffic classification irrespective of its smaller training time for all ML algorithms as compared to that of Correlation based FS algorithm. So, from these results it is obvious that Correlation based FS algorithm is a better choice for feature reduction in order to make IP traffic classification real time compatible.

Table 2: Classification Accuracy and Training Time of five ML Classifiers for Dataset $2_{\text{correlation}}$

ML Classifiers	MLP	RBF	C4.5	Bayes Net	Na ïve Bayes
Classification Accuracy (%)	46.25	80.625	83.125	91.875	84.375
Training Time (Seconds)	4.47	7.38	0.36	0.36	0.08

Table 3: Classification Accuracy and Training Time of five ML Classifiers for Dataset $2_{\text{consistency}}$

ML Classifiers	MLP	RBF	C4.5	Bayes Net	Na ïve Bayes
Classification Accuracy (%)	34.375	75	81.875	85	70
Training Time (Seconds)	3.97	35.98	0.31	0.06	0.01

Table II and figure 5 shows that maximum classification accuracy is provided by Bayes Net classifier i.e. 91.875 % which is much larger than that in case of Dataset 1 which is full feature dataset captured at 2 second packet capture duration. This is because of elimination of irrelevant and redundant features which are the main cause of miss-classification. Training time of Bayes Net Classifier is also appropriate i.e. reduced from 0.7 Seconds to 0.36 Seconds only in Dataset $2_{\text{Correlation}}$.

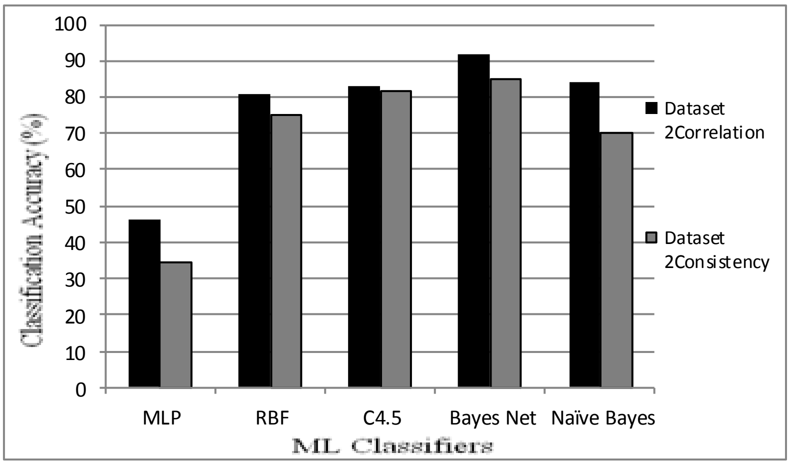

Fig. 8: Classification Accuracy of five ML Classifiers for Dataset $2_{\text{Correlation}}$ and Dataset $2_{\text{Consistency}}$

Figure 9 and 10 show Recall and Precision values of individual internet applications for three classifiers whose performance is consistent for all the datasets i.e. C4.5, Bayes Net and Na ïve Bayes. It is evident that Bayes Net gives 100% recall value for IM, VoIP and Software Update applications and 100 % precision values for P2P, IM VoIP and SU applications in case of Dataset $2_{\text{Correlation}}$.

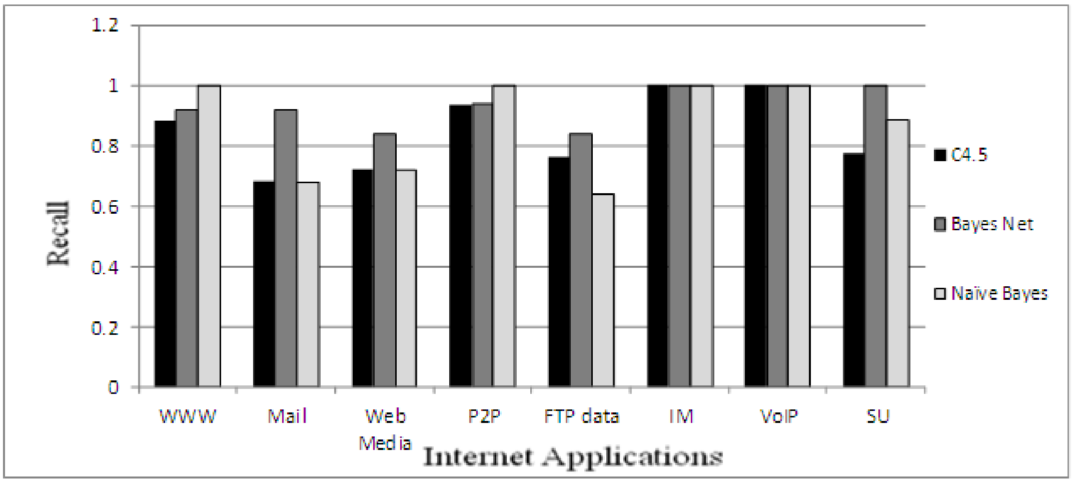

Fig. 9: Comparison of Recall for Dataset $2_{\text{Correlation}}$

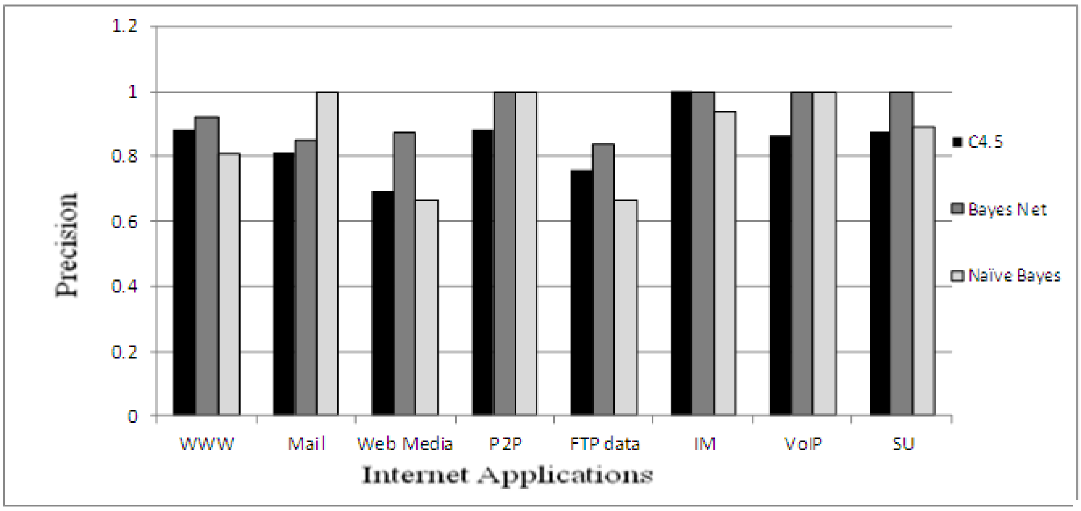

Fig. 10: Comparison of Precision for Dataset $2_{\text{Correlation}}$

From all this discussion and analysis of 2 second duration full feature and reduced feature datasets, it is evident that Bayes Net algorithm gives better performance in terms of classification accuracy and training time. For this, packet capture duration should be as small as possible and number of features used to characterize each application sample should also be reduced to much extent using Correlation based Feature Selection Algorithm. Thus IP traffic classification becomes more real time compatible and online using Bayes Net classifier.

VI. Conclusions and Future Scope

In this paper, firstly real time internet traffic has been captured using Wireshark software for packet capture durations of 2 seconds. After that, Internet traffic from this dataset is classified using five ML classifiers. Results show that Bayes Net Classifier gives better performance with classification accuracy of 88.125%. But the problem with this technique is large training time which makes it ineffective of real time and online IP traffic classification. Solution of this problem is reduction in number of features characterizing each internet application sample. For this Correlation based FS algorithm is better choice with which a reduced feature dataset has been developed. Using this new dataset, performance of five ML classifiers has been analyzed. Results show that Bayes Net classifier gives better performance among all other classifiers in terms classification accuracy of 91.875 %, training time of ML algorithms and recall and precision values of individual internet applications. Thus it is evident that Bayes Net is an effective ML techniques for near real time and online IP traffic classification with reduction in packet capturing time and reduction in number of features characterizing application samples with Correlation based FS algorithm.

In this research work, the packet capturing duration is reduced to 2 seconds to make this approach suitable for implementing real time IP traffic classification. For this purpose, the packet capturing duration should be as less as possible. This can be further reduced to fraction of seconds which will make this classification technique more real time compatible. Secondly, this internet traffic dataset can be extended for many other internet applications which internet users use in their day to day life and it can also be captured from various different real time environments such as university or college campus, offices, home environments and other work stations etc.

References

[1] Thuy T.T. Nguyen and Grenville Armitage. A Survey of Techniques for IP 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] http://www.iana.org/assignments /port-numbers

[4] Abuagla Babiker Mohd and Sulaiman bin Mohd Nor. Towards a Flow-based IP traffic classification for Bandwidth Optimization, International Journal of Computer Science and Security (IJCSS), Vol. 3, Issue 2, pp. 146-153.

[5] A.W.Moore and D.papagiannaki, Toward the accurate Identification of network applications, in poc. 6th passive active measurement. Workshop (PAM), mar 2005, Vol.3431, pp 41-54.

[6] Subhabrata Sen, Oliver Spatscheck, Dongmei Wang. Accurate, scalable in-network identification of p2p traffic using application signatures, in Proceedings of the 13th international conference on World Wide Web, New York, NY, USA, ACM: pp: 512 – 521.

[7] Thomas Karagiannis, Konstantina Papagiannaki and Michalis Faloutsos, BLINC: Multilevel Traffic Classification in the Dark, in SIGCOMM'05, August 21–26, 2005, Philadelphia, Pennsylvania, USA.

[8] 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.

[9] Li Jun, Zhang Shunyil, Lu Yanqing, Zhang Zailong. IP traffic classification Using Machine Learning, Nanjing University of Posts and Telecommunications, Nanjing 210003, China.

[10] 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.

[11] Kuldeep Singh and Sunil Agrawal. Comparative Analysis of five Machine Learning Algorithms for IP Traffic Classification, In International Conference on Emerging Trends in Networks and Computer Communications (ENCTT-2011), Udaipur, Rajasthan, India, April 22-24, 2011.

[12] Mark A. Hall, Correlation-based Feature Selection for Machine Learning, University of Waikato, Hamilton, New Zealand, April, 1999.

[13] Manoranjan Dash, Huan Lau, Consistency – based search in feature selection, Artificial Intelligence, Elsevier, 27 March, 2003.

[14] 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.

[15] 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.

[16] 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.

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

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

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

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

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

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

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

Kuldeep Singh received his B. Tech degree in Electronics & Communication Engineering in 2009 from Punjab Technical University, Jalandhar, Punjab, India and M.E. degree in Electronics & Communication Engineering in 2011 from Panjab University, Chandigarh, India. He is Assistant Professor at RIMT Maharaja Aggrasen Engineering College, Mandi Gobindgarh, Punjab, India. He has published several research papers in national and international conferences and journals. His research area includes Machine Learning techniques, wireless sensor networks.

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 19 years of teaching experience (undergraduate and postgraduate classes of engineering) and has supervised several research works at masters level. He has several 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.

B. S. Sohi received his B Sc. Engineering Degree in Electronics and Electrical Communication Engineering in 1971, M.E. degree in Electronics Engineering in 1981 and Ph. D degree in Electronics Engineering in 1992 from Panjab University, Chandigarh, India. He is campus director of Chandigarh Group of Colleges, Gharuan, Punjab, India. He has 35 years long teaching, administration and R & D experience and has supervised several; research works at doctorate and masters level.. He has 105 research papers to his credit in national and international journals and conferences. The author’s main interests include wireless communication, networking, applications of artificial intelligence etc.

How to cite this paper: Kuldeep Singh, S. Agrawal, B.S. Sohi,"A Near Real-time IP Traffic Classification Using Machine Learning", International Journal of Intelligent Systems and Applications(IJISA), vol.5, no.3, pp.83-93, 2013.DOI: 10.5815/ijisa.2013.03.09