Sindh Univ. Res. Jour. (Sci. Ser.) Vol.45 (4) 731-736 (2013)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Machine Learning Based Intelligent System for IP Traffic Classification

M. SOHAIB, A. KHALIL*, K. AHMAD*, J. KHAN*, F. IMTIAZ*

Department of Computer Science and Information Technology, Superior University, Lahore

Abstract: As the world has become a global village, internet is considered as preliminary component of each institution, corporation, business as well as an individual. As internet base applications varies on the basis of functionality they provide to the end users, their bandwidth requirements also varies on the basis of Quality of Service (QoS) they provide. Due to the new trend introduced for classification in internet applications like use of dynamic port numbers i.e. P2P applications, classical classification techniques i.e. port based and pay-load based classification are considered inefficient. So to minimize the risk of inefficient IP data classification researchers have adopted Machine Learning (ML) based IP data classification methods. In this paper an effort is made by introducing different classifiers to increase the efficiency of existing Machine Learning techniques for IP data classification.

Keywords*:* IP traffic classification, classification model, Machine Learning Algorithms, ML Classifiers, Machine learning

1. INTRODUCTION

In the recent years internet users have been tremendously increased in number, due to the increased usage of internet for carrying different tasks smoothly numerous types of internet applications have been developed in last decades. These applications consist of bulk data transfer, interactive, collaborative and also real time applications. Each pre-mentioned category of application needs different Quality of Service (QoS), due to which bandwidth demand vary for each of it. These applications demand for diverge Quality of Service (QoS) (Y. Bernet *et al.* 1999, S. Blake *et al.* 1998) and due to this fact, the bandwidth requirements for these applications also vary. For the efficient utilization of channel's bandwidth different QoS mechanisms are being suggested by the researchers but the QoS-guaranteed applications development is still bottle neck because of the inefficient IP data classification.

Earlier IP data classification algorithm i.e. port based IP data classification criteria, uses standard ports i.e. TCP and UDP ports for the transfer of IP data. Now a days this technique is considered as inefficient for classification purposes because of the usage of dynamic ports in IP network applications. Later on researchers introduced DPI technique, but did not get popular among the research community due to observed threats to legal privacy and wide spread encryption.

Letter on introduction of Machine Learning algorithms for IP data classification was highly appreciated. Classification carried out using ML algorithms provides relatively higher precision and accuracy. ML based IP data classification passes through number of steps to achieve the desired results. First of all important features are extracted from the data set and then ML-based classification model is constructed. At the end using the associated features of a known class the constructed classification model is trained.

The purpose of this research work is to study and applying different ML classification algorithms to IP data to achieve better classification accuracy and precision. Here effect of discretization and resampling to the data has been observed, which resulted in less computation time for classification process. That is pleasant increment to real time IP data classification system development.

Organization of the paper is as follows. In section II, detail Literature review is presented regarding Internet Protocol (IP) traffic classification. In section III, an overview about the classification algorithms used in this work is given. Sections IV and V are about discussion regarding the experiments and results. While in section VI is about the conclusion made from this research work.

2. RELATED WORK

During the last few decades Machine Learning approaches for IP traffic data classification have attracted the Researcher's concern (Moore *et al.* 2005, Roughan, *et al.* 2004). (Moore *et al.* 2005) collected a set of data which consists of objects having set of features and made available to the research community for further research purposes. Data collection was carried out from high performance IP networks. The purpose of the research work was to make data availability for the study of probabilistic classification techniques. Further for data characterization purposes various discriminators derivation was also the part of

++Corresponding Author: M. Sohaib, suhaib_durrani@yahoo.com, Cell: 0333-9952744

*Department of Computer Systems Engineering, University of Engineering and Technology Peshawar, Pakistan

this research work. (William *et al.* 2006) elaborated that how to check the ML algorithms performance and also described the classification accuracy of the IP data by the selected classifiers. Four different types of ML algorithms were selected for classification purpose in this work. (Donghong *et al.* 2004) carried out work to categorize IP traffic into different groups on the basis of applications. He categorizes the IP traffic into four different classes i.e. Bulk Data Transfer, Interactive, Multimedia and on the basis of QoS requirements. Feature set optimization techniques were also applied in this work to improve the classification performance. At the end different classifiers were applied to the optimized set of data to check the classification performance of the classifiers. Conclusion was derived that Decision Tree performs better than other classifiers.

In (Qin. *et al.* 2011, Dietterich, 1997) concise review of the machine learning algorithm type is presented. ML algorithms are divided into two main categories on the basis of data they used for classification purposes.

In this research work we made effort to find ML algorithms for better IP traffic classification. Moreover Discretization and resampling techniques made classification accuracy even better.

3. IP CLASSIFICATION MODEL

Machine Learning Algorithms discussed in (Qin. d. *et al.* 2011, T. Lim *et al.* 2000) provided better contribution in IP data classification. Broadly ML algorithms can be divided into two categories i.e. (i) supervised, (ii) unsupervised. In supervised learning we consider data as set of two input variables {X, Y}, where X denotes the data elements and Y presents the labels to the input elements. The main purpose of the supervised learning is to confine label, let suppose Y for the input data elements, in this case it is denoted by X for better classification results. ON the hand in unsupervised learning no labels are used to represent the input data. In this method classification is carried out just on the basis of input data elements in X. to achieve better and efficient classification we have selected supervised algorithms. In this paper we have performed different experiments on input IP data. Out of the different experiments here results of the 6 supervised classification algorithms are presented. These algorithms are discussed one by one as under.

NAÏVE BAYES

Updateable is the updated version of basic naïve bayes classifier and is placed in the supervised learning category of the classifiers. Some new updates like dealing with missing values, nominal classes are made to the basic bayes classifier in this version. The core mathematical for of naïve bayes classifier is as fallow.

$$P[C_i \backslash D] = (P[D \backslash C_i])/P[D] \qquad (1)$$

Where C_i is the class and D is the document.

4. MATERIAL AND METHODS

Multilayer Perceptrons

It is basically an artificial neural network (ANN) which approximately classifies the input data. It uses network of artificial neurons of unknown instances for feed forward and back propagation mechanisms. To every neuron a weighted input is provided, which is then manipulated by a nonlinear activation function to provide neurons output.

Feed Forward Mechanism Neural network could be considered as the ultimate implementation of human mind. However, constraints in the technology confine us to settle for much simpler design. The best solution is to design small scale electronic circuits having transfer function that of biological neurons. But this implementation is still very hectic and impractical. To overcome this complexity a layered approach is used to design neural network in which each neutron in on layer is connected to the next layer neuron. And the signal propagation is allowed in only one direction. This kind of implementation is known as feed forward neural network and is practical to build.

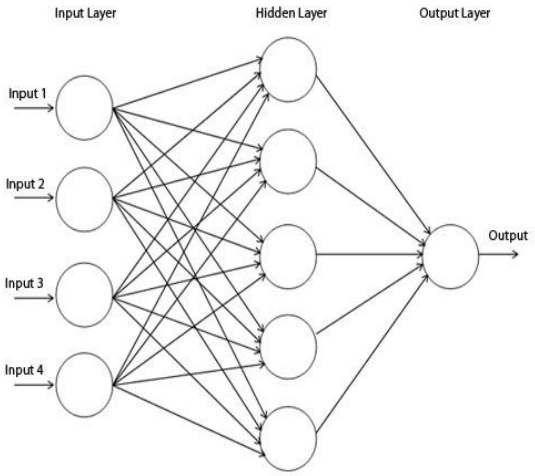


Fig. 1: Demonstration of Feed-Forward Neural Network

Referring to **Figure 1,** Feed-Forward Neural Network is shown. In this network each neuron receives input from its previous layer. The received input is then multiplied with a unique weight and is then summed and passed through the activation function which limits the output in fixed range of values.

A feed-back network is not limited to one direction. It traverses in all directions, introducing feed-back mechanism in neural network. It helps in solving even more complex problems and is open area of research.

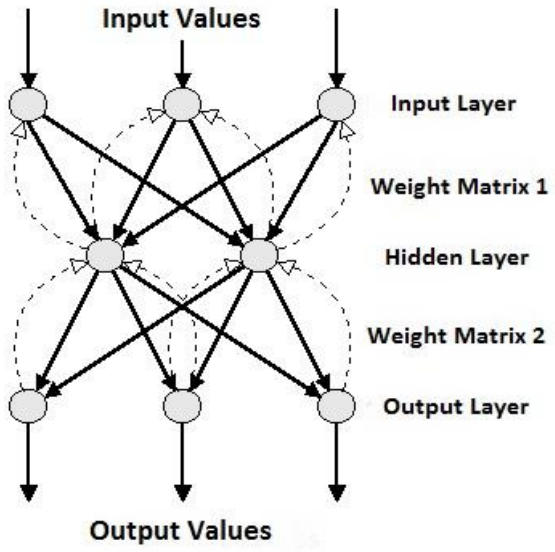


Fig. 2: Back propagation of Multilayer Perceptron

In **figure 2** feed-back network of neuron is shown. In this beside forward propagation mechanism backward propagation is also allowed.

JRip

JRip implements Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was introduced by William W. The basic of this technique is reduced error pruning (REP) based on association rules. This technique is used in decision tree classifier. In REP training data is divided into two parts i.e. growing set and pruning set. In the start formation of initial rule set is carried out which is done by over testing the growing set by some heuristic method. Then simplification method is carried out by applying one of the pruning set operators. The ultimate goal of the pruning operator application is the reduction of errors. The pruning step ends if the pruning operator applied results in the increment of errors.

J48

J48 is basically the generation of pruned or UN pruned C4.5 decision tree. C4.5 is the algorithms developed by Ross Quinlan for the generation of statistical decision tree for classification purposes. C4.5 like ID3 develops a tree from input data, on the basis of Entropy concept. The training data set consists of pre classified data I= {I1, I2, I3…..}. Each sample is in vector form i.e. Ii= {f1, f2, f3…}, where f1, f2, f3… represents different features of a sample. The training data set contain additional information with itself. The additional information consists of a class information vector, i.e. O= {O1, O2, O3….}, where O1, O2, O3 denotes different classes to which samples belong. At every node decision tree selects an effective attribute of the input data on the basis of which splits the data into different classes.

RANDOM FOREST

Random Forest is another statistical ensemble learning algorithm used for classification of input data. The need of Random Forest occurs where Decision Tree fails to give better results due to some limitations i.e. low prediction accuracy in huge dimensionality sample space and high variance. Therefore Random forest technique is developed on the basses of bagging. Random forest algorithm is as fallows.

- Bootstrap samples are formed for n-trees drawn from the input data set.
- By combining learners in parallel bootstrap aggregation is done.
- At the end prediction is carried out by mean prediction (regression) or by classification.

RANDOM TREES

Random trees classifier is also a statistical classification algorithm and was introduced by Leo Breiman. It is based on both no pruning classification and regression criteria. It takes the input data samples and classifies these samples with each instance of tree. The out of the algorithms is class label having majority votes. In case of regression problems average response of the tree is noted.

ADAPTIVE BOOSTING

Adaptive boosting (AdaBoost) is a meta-algorithm of machine learning, developed by and can be accompanied with other learning algorithms to improve classification results. It is considered adaptive in the sense that it correctly classify samples correctly that are misclassified by many algorithms. It is very sensitive to subjected noise. Another positive aspect of AdaBoost is that, it is less susceptible to overfiting problems. AdaBoost develops and invokes week classifier in each turn of round. In its each call weight's distribution is updated. As a result of this incorrectly classified samples get higher weight values on the other hand correctly classified samples get low valued weights.

5. DATA SET AND FEATURES SET SELECTION

In this work Data set regarding IP traffic is obtained from publically available online BRASIL (Lim *et al.* 2000) database for our experiments. The IP data was capture during different time intervals on

different days. the data used in this work was captured in on different days. The selected data sets consist of twelve features for each sample data was optimized to achieve optimized data sets discussed in (Blake 1998). Data set is broadly divided into four groups on the basis of different tasks performed by IP network applications. These four categories are shown in the following **Table-1.**

Table I: Categories of IP Network Applications

Class	Delay	Interactive	Bandwidth	Reliability
Inter	High	Many	Low	Mid
Bulk	-	Less	High	High
Multimedia	High	Mid	Mid	Mid
Service	Mid	Mid	Low	High

Features set selection is the most important and challenging task in task in Machine Learning. The selected feature set must be able to classify input data samples accurately. Researchers have contributed a lot in the IP data feature set extraction and selection. Instead of beginning from the scratch in this paper already extracted and optimized feature set has been used for experiments. They called the modified feature set as FRS. Features selected in FRS feature set are **Table-2** given below.

Table 2: FRS Features Set

Abbreviation	Label	Description
Max_data_ip_ab165	C	Max: total bytes in an IP packet
Var_data_ip_ba187	E	Variance of the total bytes in an IP packet
Min_segm_size_ab83	D	The min: segment size observed, throughout the connection. (client-> server)
Min_segm_size_ab84	B	The min: segment size observed, throughout the connection (server-> client)
Prtcl	A	Protocol number

6. EXPERIMENTATION AND RESULTS

Computing performance metric

The degree of the performance of a classifier is calculated on the basis of accuracy precision and recall metrics. These are defined as under.

- Accuracy: It is the percentage of correctly classified samples of input data.
- Recall: Percentage of sample members of class X correctly classified as belonging to class X.
- Precision: Percentage of those samples that are really the members of class X, among all those samples which are classified as class X.

For the comparison and verification of adopted ML-algorithms k-fold cross validation criteria is used. At every round one subset out of k-1 is selected for testing the developed model and rest are used for training purpose. The whole process is carried out k times.

7. RESULTS AND DISCUSSIONS

After the implementation of discretization and re-sampling to the selected data set, number of instances per set optimized to its best level. For this reason k-fold the adopted cross-validation is the adopted to evaluate the performance of selected classification algorithms on the optimized data set. The k-fold value is set to be 10 for this experiment. TABLE III shows the performance of applied ML-algorithms to the optimized data set.

In **Table 3** accuracy of different ML-algorithms on the optimized data set has been shown. From this it is obvious that the decision trees classifier has perform the same on the same data set of FRS. Its accuracy has been consistent (). Interestingly, Random Forest and Random Trees have greater performance than that of Decision

Table 3: Performance of the Applied Classifiers.

Classifier	Accuracy (%)	Precision	Recall
Random Forest	98.40 ±1.5	100, 98, 100, 97	96,98,100,100
Random Trees	97.56 ± 2.00	100, 95, 98, 100	97, 100, 94, 91
Decision Tree	96.50±3.91	1 0 0 , 9 1 , 1 0 0 ,9 6	90, 100, 96, 100
\ Naïve Bayse Updateable	94.00 ± 5.13	95, 94, 86, 93	90, 85, 98, 96
Adaptive Boosting	89.00 ± 6.00	90, 89, 84, 89	88, 86, 84, 81
Multilayer perceptron	76.75 ± 4.66	75, 74, 84,80	71, 78, 84, 73
JRip	71.00± 5.6	70, 71, 79, 85	70, 69, 72, 88

Trees. Their accuracy is (98.40) and (97.56) respectively. Naïve Bayes Updateable and Adaptive Boosting also performed well but was not better than that of Random Forest and Random tress. Their accuracy have has been recorded (94.00 %) and (89.00) respectively. Accuracy for Multilayer Perceptron is 76.75 and that of JRip is 71.00 %. From the close observation accuracy of Random Forest is heights among all the classifiers.

8 CONCLUSION

Due to the increased usage of internet all over the world, there is immense need of developing new IP network application to meet the users need and also there is need to develop better IP classification techniques. The available IP network applications can be divided broadly in four main categories i.e. Multimedia, Interactive, Bulk data transfer and services. To classify these applications data correctly researchers have introduced ML- classification algorithms for IP traffic classification. Due to the introduction of such algorithms in IP traffic classification, the classification process became efficient. In this work some classification algorithms are run on the IP traffic data and their accuracy has been noted. According to different experiment carried out in beside classification data discretization and resampling is also carried out in this work. This helped in improving the classifiers performance. As a future prospective of this work hybrid classifiers can be developed for IP data classification task, which will result in even improved performance.

REFERENCES:

Blake, S., S. Shulzrinne. D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss. (1998) An Architecture for Differentiated Service. RFC 2475, IETF.

Dietterich, T. G., G. Bakiri. (1995) Machine Learning. Journal of Artificial Intelligence, Volume 2, 263-268.

Dietterich. T. G. (1997) Machine Learning Research: Four Direction. Al Magazine, volume 18, No. 4, 97-136.

Jorden. M. (2004) Learning in Graphical Models. 1st MIT Press Ed. Cambridge, MA; London: MIT Press; 1999.

Lim, T., W. Loh, Y. Shih. (2000) A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms. Machine Learning, Kluwer Academic Publishers, Boston, (2000) volume 40, 203-229.

Moore A. and D. Zuey. (2005) Internet traffic classification using bayesian analysis techniques. Proceedings Segmetrics '05': International Conference on Measurement and Modeling of Computer Systems, 50-60.

Qin, D., J. Yang, J. Wang, B. Zhang. (2011) IP Traffic Classification Based on Machine Learning. IEEE 13th International Conference on Communication Technology (2011), 882-886.

Roughan, M., S. Sen, O. Spatscheck, N. Duffield. (2004) Class-of-service mapping for qos: a statistical signature based approach to IP traffic classification. Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, New York (2004), 135-148.

Williams, N., S. Zander, G. Armitage. (2006) A Preliminary Performance Comparison of five Machine Learning Algorithms for Practical IP traffic flow classification. ACM SIGCOMM Computer Communication Review, Volume 36 Issue 5, October (2006) 5-16.