Internet Traffic Classification Using supervised Learning Algorithms – A Survey

T.Kalaiselvi[1], P.Shanmugaraja[2]
[1]PG Scholar/Department of IT, Sona College of Technology, Tamilnadu, India
[1]kalaiselvi913@gmail.com
[2]Associate Professor/ Department of IT, Sona College of Technology, Tamilnadu, India
[2]shanmugarajap@gmail.com

***

Abstract:

Growth of internet has been increased so nowadays millions of users are using internet. So managing and controlling the network became an important task. By classifying the traffic we can provide network security by blocking unwanted traffic. Traffic can be classified using port-based, payload-based and machine learning techniques. But there are some drawbacks in port-based and payload-based approaches so machine learning based classification is preferred. In machine learning techniques features about the flows like inter packet arrival time, size of the flow, etc., are used. Each flow uses the same set of features but the value of the feature varies. Selecting the feature is an important task for improving the performance and as well as accuracy. There are two types of machine learning techniques. They are supervised and unsupervised. Here we are going to present a survey on supervised learning algorithms along with different feature sets.

Keywords: Traffic Classification, machine learning, feature selection

1. INTRODUCTION

Traffic classification through the network consists of flows from different applications. Some of the applications are sensitive to delay whereas others are insensitive to delay. So that the requirements vary for each and every applications. So traffic has to be classified.

Traffic classification is the process of identifying and classifying the protocols or applications available in the network. Traffic classification is beneficiary for network management and providing QOS support, network security. With traffic classification we can identify the insights of traffic.

Traffic classification schemes are port-based, payload-based and statistical-based. Port-based scheme examines the port number which is available in the header of the packet. Due to the advent of dynamic port assignment and some of the port numbers are not registered in IANA (Internet Assigned Numbers Authority) port-based scheme is not preferred. Payload-based scheme [6] inspects the payload. Due to the privacy regulation and the advent of cryptographic techniques payload-based scheme is not preferred.

Since the traditional schemes became ineffective, statistical-based methods are preferred by researchers. There are two types of machine learning techniques [2]. They are supervised and unsupervised learning. Supervised learning (classification) classifies the traffic based on the labelled data. But obtaining the labelled data is a difficult process. Whereas unsupervised learning (clustering) divides the unlabeled data based on the similarity among the data.

For classifying the traffic the following clustering algorithms can be used k-means, Expectation Maximization (EM), DBSCAN. Supervised machine learning algorithms are decision tree (c4.5, Random forest), naïve Bayes. [7] States that identifying the traffic features is important for classifying the traffic. But identifying the feature is more important than selecting the algorithm. The traffic features are calculated from the flow. Features are the statistical characteristics calculated from numerous information of objects such as mean packet length, packet arrival time, total flow length, etc.,

2. PERFORMANCE METRICS

Five metrics [5] are used for evaluating the performance of machine learning algorithms. They are overall accuracy, precision, recall, F-measure and classification speed. Metrics are given in the following.

Overall accuracy: It is the ratio of number of traffic flows classified correctly to the total number of flows available in the dataset.

Precision: Percentage of flows which are correctly attributed to the given application.

Recall: Percentage of flows available in an application class which are identified properly.

F-measure: By using the harmonic mean of precision and recall we are calculating F-measure.

2×precision×recall÷(precision+recal)

Classification speed: Number of classification decisions taken per second.

3. TRAFFIC CLASSIFICATION USING MACHINELEARNING ALGORITHMS

By using machine learning large number of flow samples can be handled. Machine learning techniques are used in various fields like medical diagnosis, predicting the load, etc. For classifying the traffic two types of machine learning algorithms are used. They are clustering and classification approach.

3.1Clustering Approach

Expectation maximization

By using the similarity among the patterns this approach classifies the traffic into different applications. It classifies traffic based on the probability distribution which has maximum likelihood on its attributes. Here each and every cluster is expressed mathematically by parametric probability distribution. [4] States Expectation Maximization algorithm and stated that their approach can achieve accuracy up to 91%. But it doesn't work for the traffic which are well known and also if the cluster size is large. It is an expansion of k-means.

K-means clustering

Main aim of k-means algorithm is to identify the minimal Euclidian distance between the defined flows and groups. The following are the steps carried out:

1) During the training phase number of the cluster centers (k) are generated randomly.
2) Each and every item is assigned the cluster center by estimating the Euclidian distance.
3) Each cluster center is mean of the items assigned.
4) Steps 2 & 3 has to be repeated until convergence.

Advantage of k-means algorithm is the computational speed will be faster if the k is small. Overall accuracy of k-means algorithm is comparatively good.

3.2Classification approach

Bayesian algorithm

In machine learning an important research direction is Bayesian learning. In Bayesian classification algorithm Bayesian theory is considered as foundation. By using priori probability Bayesian theorem computes posterior probability.

When traffic is classified using Bayesian classification algorithm then we can achieve 65% accuracy when 248 features are used. [1] Represented two improvement works: naïve Bayes kernel density and fast correlation based filter. After this work the accuracy achieved is 96%.

4. COMPARISON OF SUPERVISED LEARNING ALGORITHMS

4.1 Discretization of features

C4.5 performs well under any situation because before classifying it discretize the input features. [7] States that the features has to be discretized as a preprocessing step before running the algorithm. Discretization is the process of transferring continuous function into discrete counterparts.

Based on the following two approaches discretization can be done

Unsupervised discretization- Quantizing each feature in the absence of any knowledge of classes of instances in training set.

Supervised discretization- It takes classes into account.

Using discretization here the features for TCP and UDP are selected. Feature for UDP packets is that the size of the first two packets contribute most in identifying UDP packets, whereas for TCP second to sixth packet do the same. If the features are discretized then the port and packet size (maximum and average) features will give better results. We can obtain the maximum and average packet size only after the completion of the flow. So early classification is not possible. The author states that classification accuracy will get missed if first few packets are missed so the flows has to be captured from the beginning.

On an average the tested algorithms Naïve Bayes, SVM, k-NN and C4.5 can achieve greater than 93% accuracy when it is dealt with the port and packet size features as discrete intervals.

4.2 Feature selection for real time traffic

In machine learning based traffic classification selecting the feature from the traffic flow plays a vital role. For identifying and selecting the best feature which yields better performance accuracy of traffic classification several feature selection algorithms are practiced. Good feature subset doesn't only improve the accuracy of the algorithms, but also improve the computational performance.

For selecting the features two feature selection algorithms are used. They are filter model and wrapper model. In filter model for determining the importance and relevance of the features, characteristics of training data is used. Wrapper model uses the result of any particular classifier on a test set for different combination of features. Here feature subsets are selected for using them in online traffic classification. So real-time feature subset is constructed.

[9] States that unsupervised algorithms like C4.5 and Random forest yields greater accuracy than others. Here only unsupervised algorithms are focused C4.5 yields greater than 96% of accuracy and random forest yield greater than 99% of accuracy in many cases. But accuracy of these decision tree based algorithms are poor while applying them for classifying P2P application. When these two decision tree based algorithms (C4.5 and Random forest) are used with real time features their build time is lowest when compared with others. So when using real time features, decision tree based algorithms yield good performance.

4.3 Feature selection algorithms

In classifying the internet traffic, selecting the features is an important aspect. So for selecting the features a hybrid method called weighted symmetrical uncertainty (WSU) method along with area under roc curve (AUC) is used. It pre-filters most of the features with WSU metric and then chooses the optimal features. After filtering most of the features with WSU metric, it is necessary to identify the optimal features to make a specific classifier achieve the best performance on the training data. These feature selection algorithms used in [3] overcomes the concept drift and class imbalance problem. Selecting the features for overcoming class imbalance and concept drift problems is very difficult process. Concept drift problem occurs due to the changes of data distribution dynamically, so for improving the performance features has to be selected. Class imbalance is the problem of classifying majority of flows alone. The features selected using this WSU and AUC algorithms are server port, minimum segment size from client to server, initial window bytes from client to server and as well as from server to client. Our selected features produce better accuracy in terms of bytes. And also C4.5 algorithm gives better performance in terms of accuracy in classification and as well as speed.

4.4 Supervised real-time traffic classification

In [8] three flow feature sets are evaluated, two of them are features generated from full flows and another one is from the early sub-flow statistics which is obtained from the first few packets of each flow. Here they have analyzed using naïve Bayes and decision tree algorithms. Naïve Bayes has achieved 90% accuracy. C4.5 achieved 97% accuracy and random forest achieved 95% accuracy. Here for the selected feature sets decision tree based algorithms performs well.

5. CONCLUSION

For understanding the evolution in traffic classification we presented this survey. In this paper, we have analyzed the accuracy of supervised learning algorithms by using different feature sets. The supervised decision tree based algorithms provides better performance and accuracy than the other supervised algorithms.

REFERENCES

[1] A. W. Moore, D. Zuev, Internet traffic classification using Bayesian analysis techniques, Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, 2005
[2]Bin Hu, Yi Shen,' Machine Learning Based Network Traffic Classification: A Survey' Journal of Information & Computational Science 9: 11 (2012) 3161–3170

[3] Hongli Zhang, Gang Lu , Mahmoud T. Qassrawi, Yu Zhang, Xiangzhan Yu, "Feature selection for optimizing traffic classification" ELSEVIER, 1457-1471,2012.

[4]J.Erman, A. Mahanti, M.Arlitt, Internet traffic classification using machine learning, Global telecommunications conference, 2006 pp.1-6.

[5] Nguyen.T.T.T, Armitage.G, 'A Survey of Techniques for Internet Traffic Classification using Machine Learning', in IEEE Communications Surveys and Tutorials, 2008.

[6] Risso.F, Baldi.M, Morandi.O, Baldini.A, Monclus.P, 'Lightweight, Payload-Based Traffic Classification: An Experimental Evaluation', 2008.

[7] Y. Lim, H. Kim, J. Jeong, C. Kim, T. Kwon, and Y. Choi, ''Internet Traffic Classification Demystified: On the Sources of the Discriminative Power,'' Proc. ACM CoNEXT, 2010, p. 9.

[8]Yu Wang and Shun-Zheng Yu, Supervised Learning Real-time Traffic Classifiers, Journal of networks, vol. 4, no. 7, september 2009 pp.622-629

[9] ZHAO Jing-jing, HUANG Xiao-hong, SUN Qiong, MA Yan, "Real-time feature selection in traffic classification", ELSEVIER, 2008.