Fine Grained Classification of Internet Video Traffics

Yu-ning DONG, Li-tao YAO
College of Telecommunications & Information Engineering
Nanjing University of Posts and Telecommunications
(NUPT), Nanjing, China
E-mail: dongyn@njupt.edu.cn; yaolitao2012@163.com

Hai-xian SHI
School of Horticulture
Nanjing Agricultural University
Nanjing, China
shx@njau.edu.cn

***Abstract*—For the purposes of efficient network resource management and QoS (Quality of Service) support of different video services, this paper proposes a fine grained classification scheme for Internet video traffics based on hierarchical clustering. We study a number of QoS and network resource requirements related statistical features of some typical video applications and validate their effectiveness in video traffic classification. This scheme classifies video services with the combinations of these statistical features. Experiments are performed on a large scale real network video traffic data. The results show that the proposed method can achieve significantly better classification performance in comparison to existing methods.**

***Keywords*—QoS; hierarchical clustering; video traffic classification; flow statistics**

I. INTRODUCTION

With the advances of network and multimedia technologies in the last decade, Internet video services grow rapidly. Meanwhile a variety of new applications make the network environment increasingly complex and a series of problems arise such as efficient network resource management, QoS (Quality of Service) guarantee of multimedia services. It is believed that accurate classification of network traffic is an effective way for Internet service providers (ISPs) and network regulators to tackle these issues [1].

Currently, there are mainly three categories of traffic classification methods: port-based classification, deep packet inspection (DPI) [2], and statistical analysis [3]. The statistics methods based on machine learning (ML) that take advantage of the statistical properties of traffic flows are attracting more attention of researchers [4-5].

Chandrakant et al. [6] implemented six ML based algorithms for Internet traffic classification and compared the performances of them using commonly used flow statistical features such as typical packet lengths and inter-arrival times. Zhang Jun et al. [7] adopted a semi-supervised ML technique to effectively discriminate zero-day application. Duo Liu et al. [8] applied Fuzzy C-means (FCM) clustering to classifying P2P (peer to peer) applications. Their work was aimed at reducing the computational complexity of FCM while keeping the clustering accurate. Zhang Meng et al. [9] proposed an encrypted traffic classification scheme based on improved FCM and the improved FCM reduces the impact of random initial clustering centers. Wang Yu et al. [10] proposed a constrained clustering scheme for Internet traffic classification, and utilized 20 statistical features to represent IP flows.

However, most previous works classified video services as one or two classes in network traffic classification. Mu Xuefeng et al. [11] studied a parallelized network traffic classification scheme based on hidden Markov model to divide video services into conversational and streaming videos. Gonçalves G.D et al. [12] only distinguished P2P video traffics. In fact, there are varieties of video services, such as streaming media videos, network live TVs, P2P-based download videos etc., and they may have different QoS and network resource requirements. For example, symmetric videos have relatively high requirements on both uplink and downlink bandwidth; network live TVs need to be provided with real-time performance; interactive video communication is sensitive to delay. In the meantime, video flows belonging to the same category generally have similar QoS/resource requirements. The correct identification of video flows can hence help ISPs to understand what level of QoS and resource requirements they need, and make appropriate resource allocation for them in order to improve the end user's experience. Therefore, it is necessary to carry out finer-grained classification for video services. There are seemingly very few works that considered this problem [13]. Still, how to find effective statistical features for video traffic classification remains a huge challenge.

In this paper, first, we study some new and more effective features for video traffic classification by extensive analysis on statistical characteristics of typical Internet video applications. Then, a new hierarchical clustering scheme is developed with proper combination of flow features for improving classification accuracy. Experimental results show that this method can achieve better classification accuracy compared with existing methods in network video traffic classification.

The rest of the paper is organized as follows. Section II presents the analysis and selection of flow statistical features. In Section III, a detailed description of a multi-layer traffic identification scheme is given. Experimental results are reported in Section IV. Finally, Section V concludes the paper.

II. FEATURE ANALYSIS AND SELECTION

Dataset: A large number of real-world Internet video traces were captured using WireShark [1] in the campus network

[1] http://wiki.wireshark.org/

environment of NUPT during the period of Oct. 2013-Apr. 2014.

By processing the raw flow data and calculating statistical features of different network video applications, we select some QoS and resource requirements related features that have better discriminative effects after extensive data analysis and mining. For this purpose, each sample flow is processed to extract over 50 statistical features, including the PDF (probability density function) / CDF (cumulative density function) / CCDF (complementary cumulative density function) of (D/U, downstream and upstream) packet size; the information entropy of (D/U) packet size; the PDF/CDF/CCDF of (D/U) packet inter-arrival time; the information entropy of (D/U) packet inter-arrival time; the ratio of downstream bytes to upstream bytes; the ratio of downstream packets to upstream packets; the number of downstream sub-flows; (D/U) packet rate; (D/U) byte rate; downstream sub-flow duration (standard deviation); the max./mean/min. value of (D/U) packet size; etc.

This work focuses on six kinds of Internet video applications: asymmetric standard definition (SD) video (e.g. Youku/YouTube[2] SD online video play), asymmetric high definition (HD) video (e.g. Youku/YouTube HD online video play), HTTP-download video data, interactive video communication class (e.g. QQ/Skype[3] video chat), P2P video data sharing (e.g. Xunlei[4] video download), and network live TV (e.g. Sopcast[5] live TV broadcasting). Here we introduce some notation concepts. A flow refers to sequences of packets captured in 30 minutes for an application. In order to study the problem, we grabbed 60 flows for each application. Total number of flows is 360, and total data amount is 13.3G bytes.

Downstream data refers to the data downloaded to local IP. Upstream data refers to the data uploaded from local IP. Previous works [1] have shown that downstream data carry more information than upstream data, so our analysis focuses on the downstream data.

In order to find more effective traffic statistical features, we select, after extensive statistical analysis, four features: ratio of downstream bytes to upstream bytes (RDBUB), information entropy of packet size downstream (IEPSD), the number of downstream sub-flows (NDSF), the number of valid downstream IP addresses (NDVIP), and their combinations to classify the typical network video applications. Below are detailed descriptions of the four features:

- Ratio of Downstream Bytes to Upstream Bytes

Ratio of downstream bytes to upstream bytes is the ratio of total received downstream data bytes to the total upstream data bytes after removal of overhead packets (e.g. control packet).

- Information Entropy of Packet Size Downstream

Information entropy of downstream packet size is defined in Equation 1. We can use it to measure the degree of uniformity of packet size distribution.

$$E = -\sum p(x_i)\log_2 p(x_i) \quad (1)$$

Where E denotes the information entropy of downstream packet size, x_i is the i^{th} packet size, and $p(x_i)$ is the probability density function of the packet size x_i.

- The Number of Downstream Sub-flows

Continuous occurrences of the same source IP packets called a sub-flow fragment; these IP addresses may repeat.

- The Number of Downstream Valid IP

It is observed that there are many interactions between local IP and remote IPs in the process of capturing data stream. Since there are some unavoidable background applications during the data capturing operation, they may bring some irrelevant IP addresses. In this case, we define valid IP as that whose continuous duration is longer than 0.5s. The number of valid downstream IPs is the total number of IPs whose continuous duration longer than 0.5s in downstream.

It is found in our statistical analysis that significant difference of features exists among different network video applications. Specifically, there is a big difference in the ratio of downstream bytes to upstream bytes between symmetric (QQ, Xunlei, Sopcast) and asymmetric traffics (ASD, AHD, HTTP-download). The ratio of downstream bytes to upstream bytes of symmetric traffic is much smaller than that of asymmetric traffic. The information entropy of downstream packet size of QQ and Xunlei are larger in symmetric traffic, while the difference between Sopcast and asymmetric traffic in IEPSD is not obvious.

As shown in Fig. 1, one can distinguish between symmetric and asymmetric applications in the two-dimensional feature space (IEPSD and RDBUB), but cannot separate these two categories using any single feature. It is easy to divide these samples into two parts with a line (see Fig. 1). This implies that the selected features combination is often more effective.

To further identify the three applications: QQ, Xunlei and Sopcast within the symmetric category, we will use the combination of the IEPSD and the logarithmic form of NDSF, as shown in Fig. 2.

For asymmetric traffics, we cannot distinguish these applications using the above three statistical features. So a new statistical feature was selected, which is the number of valid IP downstream. HTTP-download is a traditional application type. This application has only interaction between two IP hosts, so the number of valid IP downstream is small. However, the captured traffic streams of asymmetric SD and HD are not from only one video server. Once the connection is down due to interference in network environment, the local IP host will establish another connection with another server. So the number of valid IP for this traffic is inevitably large. We can distinguish HTTP-download from the other two applications

[2]http://www.youtube.com/
[3]http://skype.gmw.cn/
[4]http://dl.xunlei.com/
[5]http://www.easetuner.com/

with this feature roughly, but certain overlap of these two categories still exists.

For the asymmetric SD and HD (ASD and AHD), from the feature space available, only the ratio of downstream bytes to upstream bytes can be used to separate them approximately. The overlap between ASD and AHD is heavy, which causes the clustering result unsatisfactory. From a realistic point of view, the definitions of ASD and AHD may not be clear and vary with the development of relevant technology. ASD and AHD traffics at different times will have different definitions. The present ASD may be similar to previous AHD five years ago, but we cannot get rid of this influence when getting the data. So this overlapping case seems to be acceptable.

Through the above analysis, we decide to use these four statistical features in the experiments.

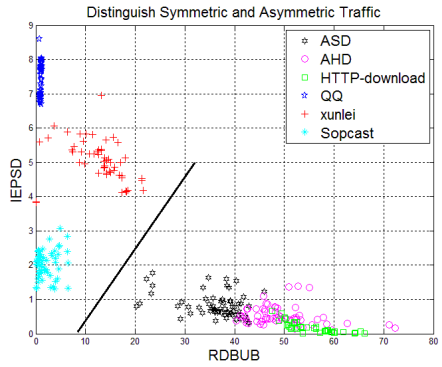


Figure 1. Distinguish between symmetric and asymmetric video applications

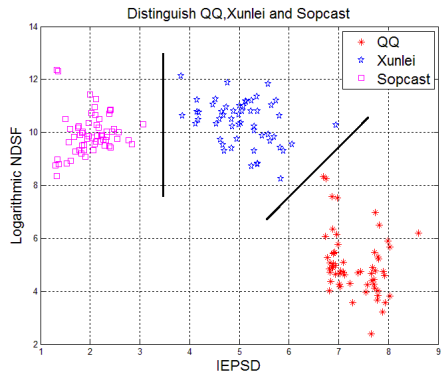


Figure 2. Distinguish among QQ, Xunlei and Sopcast

III. HIERARCHICAL CLUSTERING SCHEME

FCM is a clustering algorithm based on objective function which attributes clustering as a nonlinear programming problem with constraints [8]. It gets fuzzy classification and clustering of data set through optimizing solution. Its basic idea is to realize the dynamic iterative clustering through revising the clustering center V and fuzzy matrix U, which makes the similarity of objects in the same cluster maximum and that in different clusters minimal.

With the increased number of applications they have to choose, the performance of all classification algorithms will deteriorate [14]. That is to say, it will be difficult to classify all kinds of traffic in one time with a large number of features.

Therefore, this paper employs a hierarchical FCM clustering scheme to classify network video services. This allows each classifier to work on a limited subset of video applications with features that are most suitable to distinguish them.

Fig. 3 shows the hierarchical clustering classification scheme we explore in this paper. Gray nodes are sub-classifiers and white nodes represent the individual video services. First, at the root, flows are split between the symmetric and asymmetric video traffic classes. We use RDBUB and IEPSD to distinguish these two categories in this stage (see Fig.1). Then, asymmetric video traffic is split into finer grained applications: ASD, AHD and HTTP-download, and symmetric video traffic is split into finer grained applications: QQ, Xunlei and Sopcast (see Fig. 2), to get the final classification results.

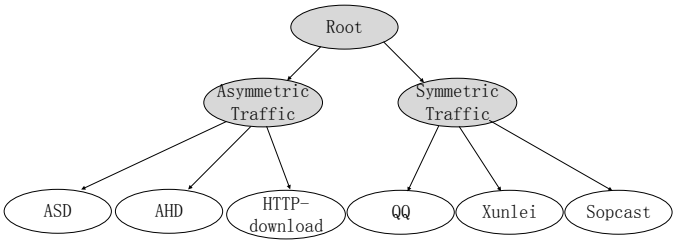


Figure 3. Hierarchical clustering scheme for fine grained video traffic classification

IV. EXPERIMENTS

A. Experiments Setup

In this section we use the hierarchical FCM to classify six kinds of network video applications, and the classification results are compared with those of previous work. The classification scheme consists of four modules: 1) network data acquisition module that captures the traffic data of the six types of network video applications; 2) raw data processing module that performs necessary preprocessing on the captured data, such as removal of overhead packets, separation of upstream and downstream data; 3) feature analysis and selection module where more than 50 flow statistical features are extracted and analyzed, and some more effective features are selected for classification purpose; 4) hierarchical clustering module that executes the classification of video traffics using the hierarchical FCM clustering algorithm.

B. Results and Analysis

As discussed above, the statistical features and feature combinations extracted in previous sections are more effective to distinguish the six kinds of network video applications. In order to further verify this, we exploit these feature combinations to identify the six kinds of network video traffics by using the proposed hierarchical clustering algorithm.

Precision (P) and recall (R) are two commonly used performance metrics of classification results [15]. They are defined as below.

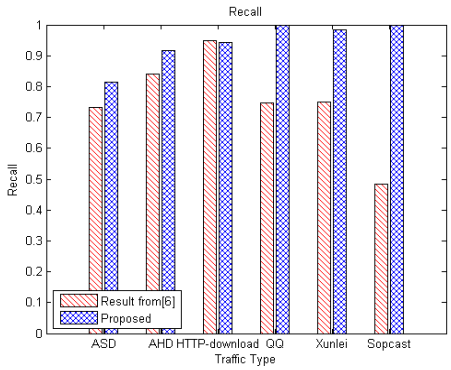
$$P = \frac{TP}{TP + FP} \qquad (2)$$

$$R = \frac{TP}{TP + FN} \quad (3)$$

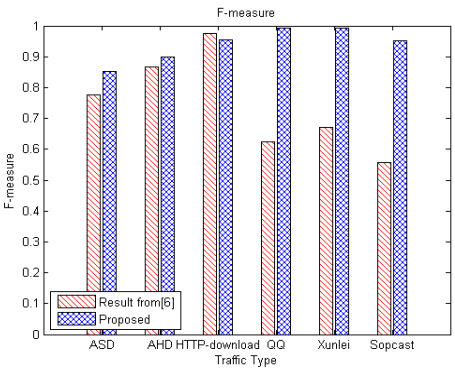
Where, for Class *j*, *TP* (true positives) is the number of correctly identified samples; *FP* (false positives) is the number of other classes' samples identified as Class *j*; *FN* (false negatives) is the number of Class *j*'s samples identified as another class. F-measure (*F*) [15] is the weighted harmonic mean of precision and recall, defined as below,

$$F = \frac{2 * P * R}{P + R} \quad (4)$$

In the following, we use the recall and F-measure metrics to validate the effectiveness of the selected features in combination with the proposed hierarchical clustering algorithm. Fig. 4 shows the recall and F-measure of different algorithms on the dataset. Result from [6] is the classification results with the method of [6] which uses the features of average packet size downstream (APSD) and average packet inter-arrival time downstream (APITD). Proposed denotes the classification results of our method using the selected feature combinations with a hierarchical FCM clustering algorithm.



(a) Recall



(b) F-measure

Figure 4. Classification results of different algorithms

As can be seen from Fig.4, all the performance metrics of our method are better than those of the method of [6], except for HTTP-download. Though the recall and F-measure of HTTP-download is lower with the proposed method, the accuracy rate is more than 95%; so the result can basically satisfy the requirements of classification for HTTP-download videos. For ASD and AHD, our method is slightly better than the method of [6]. As depicted in Fig.4, both recall and F-measure of ASD and AHD are less than 90% for both methods, which indicates that there is no obvious boundary between ASD and AHD. However, both performance metrics of our method for QQ, Xunlei and Sopcast are much better than [6]. The reason is that according to the analysis in Section III, we use different statistical features to classify different video streams, while method [6] used same features to identify all the video streams (see Fig.5). As shown in Fig. 2, we can properly distinguish QQ, Xunlei and Sopcast in the two-dimensional feature space, while the overlap among them is serious in Fig.5. So the performance of our method for these three video applications is much better than the method of [6].

Fig.5 shows that the boundary between HTTP-download and AHD is more obvious with the features of APSD and APITD, so the classification results for HTTP-download are slightly better than ours. By comparing Fig. 5 with our method one can see that there are overlaps between ASD and AHD, which makes the performance of these two methods similar. On the whole, the proposed algorithm has obvious advantages over the method of [6].

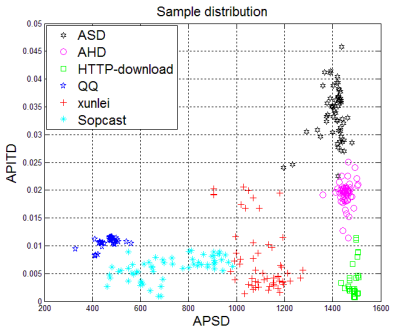


Figure 5. Sample distribution of method[6] using features of average packet size and packet inter-arrival time downstream

V. CONCLUSION

This paper presents a hierarchical clustering classification algorithm based on new flow statistical feature combinations. We demonstrate the effectiveness of these feature combinations by experiment with real network traffic traces. The proposed algorithm using these feature combinations has better overall classification results than the method of [6]. Although each layer using different combinations of features in the hierarchical clustering algorithm incurs a certain degree of complexity, considering the improved classification accuracy the increased complexity seems to be acceptable.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (No.61271233, 60972038), and the Specialized Research Foundation for the Doctoral Program of Higher Education of China (No.20103223110001).

REFERENCES

[1] Dainotti, Alberto, Antonio Pescape, and Kimberly C. Claffy. "Issues and future directions in traffic classification." Network, IEEE 26.1 (2012): 35-40.

[2] Takeshita K, Kurosawa T, Tsujino M, et al. "Evaluation of HTTP video classification method using flow group information."

Telecommunications Network Strategy and Planning Symposium (NETWORKS), 14th International. IEEE, 2010: 1-6.

[3] DONG Yuning, WANG Zaijian, FANG Shuguang, ZHANG Jian. "Survey of Methods for Traffic Identification and Classification in Multimedia Communications."Journal of Nanjing University of Posts and Telecommunications (Natural Sciences), 2013, 33 (3) : 35-44 (in Chinese).

[4] Raahemi B, Zhong W, Liu J. "Peer-to-peer traffic identification by mining IP layer data streams using concept-adapting very fast decision tree."Tools with Artificial Intelligence. ICTAI'08. 20th IEEE International Conference on, 2008, 1: 525-532.

[5] Li W, Moore A W. "A machine learning approach for efficient traffic classification." [C]//Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, 2007. MASCOTS'07. 15th International Symposium on, IEEE, 2007: 310-317.

[6] Chandrakant J R, et al. Machine learning based Internet traffic recognition with statistical approach. Proc. 2013 Annual IEEE India Conference (INDICON), Mumbai, India, Dec 13-15, 2013: 1-6.

[7] Zhang J, Chen X, Xiang Y, Zhou J, Wu J. "Robust Network Traffic Classification." IEEE/ACM Transactions on Networking, May 2014, DOI: 10.1109/TNET.2014.2320577.

[8] Liu D, Lung C H. "P2P traffic identification and optimization using fuzzy c-means clustering."Fuzzy Systems (FUZZ), 2011 IEEE International Conference on. 2011: 2245-2252.

[9] Zhang M, Zhang H, Zhang B, Lu G. Encrypted traffic classification based on an improved clustering algorithm,Trustworthy Computing and Services. Springer Berlin Heidelberg, 2013: 124-131.

[10] Wang Y, Xiang Y, Zhang J, et al. "Internet Traffic Classification Using Constrained Clustering." IEEE Transactions on Parallel and Distributed Systems, 2014, 25(11), pp. 2932-2943.

[11] Mu X, Wu W. "A Parallelized Network Traffic Classification Based on Hidden Markov Model."Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2011 International Conference on. IEEE, 2011: 107-112.

[12] Gonçalves G D, Cunha Í, Vieira A B, et al. "Predicting the level of cooperation in a Peer-to-Peer live streaming application." Multimedia Systems, 2014:1-20.

[13] Elnaka A M, Mahmoud Q H. Real-time traffic classification for unified communication networks. Proc. 2013 Int Conference on Selected Topics in Mobile and Wireless Networking (MoWNeT), 2013: 1-6.

[14] Grimaudo L, Mellia M, Baralis E. "Hierarchical learning for fine grained internet traffic classification."Wireless Communications and Mobile Computing Conference (IWCMC), 2012 8th International. IEEE, 2012: 463-468.

[15] Zhibin Y, Kil G B, Kim S. "Traffic classification based on visualization."Networked Embedded Systems for Enterprise Applications (NESEA), 2011 IEEE 2nd International Conference on. IEEE, 2011: 1-6.