

Fine Grained Classification of Internet Multimedia Traffics

Yu-ning DONG, Kai WANG

College of Telecommunications & Information Engineering
Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China
E-mail: dongyn@njupt.edu.cn; stefanmark@163.com

Abstract—For the purposes of efficient network resource management and QoS (Quality of Service) support of different multimedia services, this paper proposes a fine grained classification scheme for Internet multimedia traffics using a novel low-complexity feature selection method based on coefficient of variation. We focus on web-browsing and network video services in this work. A number of QoS and network resource requirements related statistical features of these applications are studied and validated by their effectiveness in web-browsing and video traffic classification. This scheme classifies multimedia services with the combinations of these statistical features. Experiments are performed on a large scale real network multimedia traffic data. The results show that the proposed method can achieve better classification performance in contrast to existing methods.

Keywords—Traffic classification; QoS; web browsing; video; coefficient of variation; flow statistics

I. INTRODUCTION

With the advances of network and multimedia technologies in the last decade, Internet multimedia 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. Classification of web application based on HTTP [8-9] has been intensely investigated in recent years. Inter-arrival times, packet sizes and packet ordering are used in [8] to identify the presence of another application tunneled using HTTP. 12 web applications including video, file sharing, audio, maps and so on are

analyzed in [9]. Web traffics are obtained by loading a large number of different websites in [10]. Then the websites are classified by a vector of average inter-arrival times and packet sizes. However, this method requires amount of repeated experiments with high time overhead. The work in [11] analyzed video, mail and social media based on HTTP protocol and defines the concept of “segment”. However, the classification result to some extent depends on the size of “segment”. In addition, the types of traffics are less, and lack of diversity. Ref.[12] presented a combination of keyword matching and statistical behavior profiles to identify audio, video and web based on HTTP.

Most previous works classify web browsing as one category for identification of web applications based on HTTP protocol. However, with the fast development of network applications, web becomes more and more sophisticated by carrying various kinds of applications. Hence, web browsing being treated as one class is no longer appropriate. Video services are also classified as one or two classes in current network traffic classification works. Mu et al. [13] studied a parallelized network traffic classification scheme based on hidden Markov model to divide video services into conversational and streaming videos. Gonçalves et al. [14] only distinguished P2P video traffics. In fact, there are varieties of web-browsing and video services, such as web-browsing without video (text and images only), web-browsing with video, streaming media videos, network live TVs, P2P-based download videos etc., and they may have different QoS and network resource requirements. For example, web-browsing without video and web-browsing with video may require different network bandwidth, 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, web-browsing and video flows belonging to the same category generally have similar QoS/resource requirements. The correct identification of web-browsing and 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 multimedia services. There are seemingly very few works that considered this problem [15]. Still, how to find effective statistical features with low time complexity algorithm for web-browsing and video traffic classification remains a huge challenge.

The main contributions of this paper are summarized as follows: 1) Fine-grained classification of web browsing is introduced for the first time from QoS and network resource management perspectives; 2) A novel low time complexity feature selection method based on coefficient of variation is proposed, which can extract more effective features than existing approaches.

The reminder of the paper is organized as follows. Section II presents a novel low time complexity feature selection algorithm. Section III analyzes the selected features. Experimental results are given in Section IV. Finally, Section V draws conclusions.

II. FEATURE SELECTION ALGORITHM

In this section, we propose a new feature selection algorithm based on coefficient of variation.

A. Basic idea

Let sample spaces P be expressed by matrix Q , each row of which on behalf of all characteristics of a sample and each column of which represents all samples of a characteristic. Effectiveness of a certain characteristic is that it is relatively concentrated within class and dispersive between classes. Coefficient of variation [16], also called coefficient of difference, reflects the degree of dispersion of datasets. The greater the coefficient of variation is, the more dispersed data is, conversely, the more concentrated the data is. Coefficient of variation (CV) is defined as the ratio of standard deviation to mean [16]:

$$CV = STD / AVG \quad (1)$$

Where, CV on behalf of coefficient of variation, STD represents the standard deviation and AVG the mean.

B. Steps of the algorithm

- 1) Discretization of each column of Q ;
- 2) Calculation of coefficient of variation IN_CV_{ij} of each characteristic of classes, where i represents class and j indicates feature;
- 3) Computation of coefficient of variation OUT_CV_j among all classes by the mean M_{ij} of each characteristic of classes;
- 4) Definition of coefficient of importance $CI_{ij} = \frac{OUT_CV_j}{IN_CV_{ij}}$;
- 5) Calculation of mean $CI_j = \text{MEAN}(CI_{ij})$ for fixed j ;
- 6) Ordering of CI_j ;

We have made theoretical analysis on the time complexity of this algorithm and prove our algorithm has lower computational complexity than those algorithms based on information gain or chi-squared. We will demonstrate the effectiveness and accuracy of our method by comparison with CFS (Correlation-based Feature Selection) [17], chi-squared [18] and information gain [19] based ones in the following sections.

III. FEATURE ANALYSIS

A. Data set

A large number of real-world multimedia traffic traces were captured using WireShark¹ in the campus network environment of NUPT during the period of Oct. 2014-Sep. 2015.

By processing the raw flow data and calculating statistical features of different network multimedia applications, we select some QoS and resource requirements related features that have better discriminative effects using the feature selection algorithm mentioned in Sec.II. The statistical characteristics include information entropy of upstream/downstream packet-size, mean/variance/min/max of packet interval-time, third central moment of packet-size and so on. A traffic flow is defined as continuous packet sequences that are produced by a certain application during a period of time. Downstream data refers to the data downloaded to local IP. Upstream data refers to the data uploaded from local IP.

This work focuses on six kinds of Internet multimedia applications (see Table 1): Skype audio, video streaming, network live TV, HTTP download, web browsing (text and images) and web browsing with video. In order to study the problem, we grabbed 60 flows for each application. Total number of flows is 360, and total data amount is 7.14G bytes.

TABLE I. DATA SET

Application type (abbreviation)	In detail
Skype audio (Skype ²)	Audio conversation
network live TV (cbox_sopcast ³)	Standard definition video (live)
video streaming (youku_iqiyi ⁴)	Standard definition video (stored)
HTTP download (http_dload ⁵)	Files or videos downloading based on HTTP
web browsing (csdn_times ⁶)	With text and images
web browsing with video (sina_NG ⁷)	With text, images and videos

B. Analysis of feature distributions

We analyze the two-dimensional distribution of some features that are extracted by our method in order to verify their effectiveness. Figs. 1 and 2 show that http_dload has the smallest packet interval-time and the biggest mean of packet-size. Therefore, HTTP download always make full use of network bandwidth and keeps the highest occupation ratio of network resources among these classes. Fig 2 shows that the mean of downstream byte-rate of cbox_sopcast and

¹<http://wiki.wireshark.org/>

²<http://skype.gmw.cn/> ³<http://cbox.cntv.cn/>;

<http://www.sopcast.cn/> ⁴<http://www.youku.com/>;

<http://www.iqiyi.com/> ⁵<http://pan.baidu.com/disk/>

⁶<http://www.csdn.net/>; <http://www.thetimes.co.uk/tto/news/>

⁷<http://www.sina.com/>; <http://news.nationalgeographic.com/>

youku_iqiyi that both belong to network video are only lower than http_dload. As for sina_NG and csdn_times, it is reasonable that their utilization ratio of network resources is below that of video streaming. Skype who is a kind of network audio has minimal demand of network bandwidth.

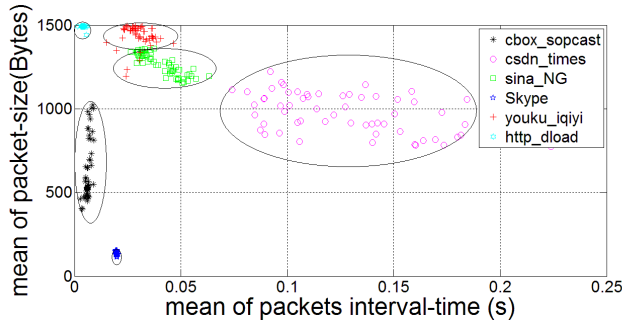


Figure 1. Two-dimensional distributions of mean of packets interval-time and packet-size

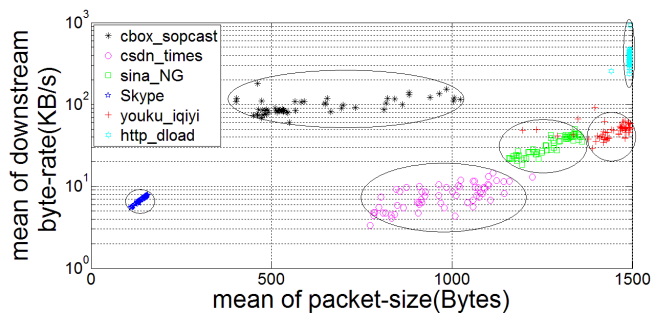


Figure 2. Two-dimensional distributions of mean of packet-size and downstream byte-rate

Further, cbox_sopcast and youku_iqiyi belongs to network live TV and video streaming, respectively. Fig 1 shows cbox_sopcast has smaller mean of packet interval-time than youku_iqiyi. One can find that the mean of packet-size of youku_iqiyi is approximately 1400 bytes, but cbox_sopcast maintains the mean of packet-size around 400-1100 bytes from Fig 2.

As for sina_NG and csdn_times, they are both web browsing, but the former contains images, text and video while there is no video in the latter. Fig 1 shows sina_NG has smaller mean of packet interval-time but greater mean of packet-size than csdn_times. As we can see from Fig 2, sina_NG has larger mean of downstream byte-rate than csdn_times. For Skype, Fig 1 shows the mean of packet-size keeps around 150-180 bytes and the mean of packets interval-time is about 0.02s. We can find that Skype's requirement of network bandwidth is the lowest among these classes in Fig 2.

From above analysis one can see that the selected characteristics are able to clearly differentiate these six kinds of applications. We will verify the validity and rationality of our method by experiment later on.

IV. EXPERIMENTS

A. Experiments Setup

In this section, we evaluate the proposed feature selection method by experiments using the six kinds of network multimedia applications as show in Table 1.

Each application contains 60 flows each of which lasts 30 minutes. Data of each application are randomly divided into training and test sets each of which contains 30 flows. This experiment will evaluate the effectiveness of our method by comparing performances of k -NN (k -Nearest Neighbor) [3] and SVM (Support Vector Machine) [4] classifier with different feature selection methods which are CFS, information gain, the proposed approach and chi-squared. Final results are the average of 20 runs.

B. Performance Metrics

To measure the performance of our approach, we use F-measure and overall accuracy, which are widely used for performance evaluation in the field of traffic classification [20] as follows:

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$F - Measure = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (4)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

Where, TP (True Positive) is the number of positive cases that were correct. FP (False Positive) is the number of negative cases that were incorrectly classified as positive. TN (True Negative) is the number of negative cases that were correct. FN (False Negative) is the number of positive cases that were incorrectly classified as negative

C. Experimental results

By using the proposed feature selection method, we obtain five features which are mean and variance of packet-size, mean of packets interval-time, ratio of downstream packets to upstream packets and mean of downstream byte-rate from 21 flow statistical features used in experiment. We get six characteristics that contain mean and variance of packet-size, mean and variance of packet interval-time, entropy of packet-size and mean of downstream byte-rate by information gain. Seven features including mean, variance and entropy of packet-size, mean of packet interval-time, mean of downstream byte-rate, ratio of downstream packets to upstream packets and ratio of downstream bytes to upstream bytes are obtained via chi-squared. 13 statistics are chosen by CFS, including mean, variance, third central moment and fourth central moment of packet-size and packets interval-time, entropy of packet-size, ratio of downstream packets to upstream packets, number of downstream sub-flow segment, fourth-order cumulant of downstream byte-rate and mean of downstream byte-rate. Figs. 3 and 4 indicate that the proposed method has higher F-measure than the other three in cbox_sopcast, csdn_times and http_dload.

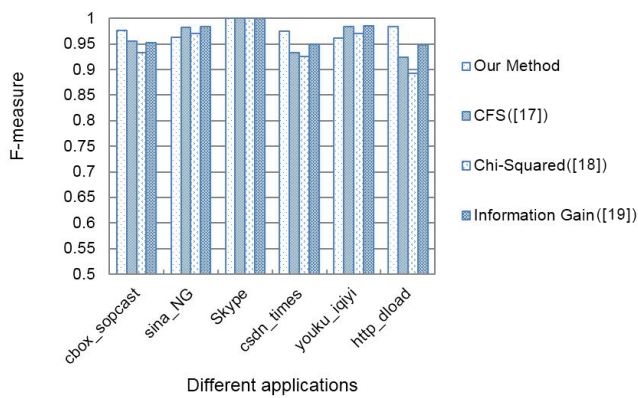


Figure 3. Comparison of F-Measure of different methods with k-NN classifier

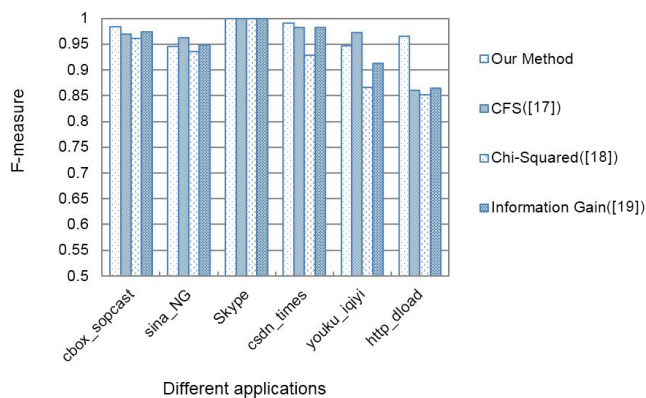


Figure 4. Comparison of F-Measure of different methods with SVM classifier

From Tables 2 and 3, the proposed feature selection method can not only effectively reduce the number of characteristics and classification time, but also improve the overall accuracy of classifiers. Our approach reaches the highest overall accuracy close to 98% among information gain whose overall accuracy is about 95% as well as CFS, and chi-squared with 94% accuracy. The results demonstrate that the proposed method can improve the classification performance in a robust way.

TABLE II. COMPARISON OF OVERALL ACCURACY OF DIFFERENT METHODS WITH K-NN CLASSIFIER

Feature Selection Method	Our method	CFS([17])	Chi-Squared([18])	Information Gain ([19])
Overall Accuracy	97.36%	94.511%	94.021%	95.012%

TABLE III. COMPARISON OF OVERALL ACCURACY OF DIFFERENT METHODS WITH SVM CLASSIFIER

Feature Selection Method	Our method	CFS([17])	Chi-Squared([18])	Information Gain ([19])
Overall Accuracy	97.36%	94.511%	94.021%	95.012%

Overall Accuracy	98.17%	95.11%	94.08%	95.78%
------------------	--------	--------	--------	--------

In general, compared with CFS algorithm in [17], our approach not only enhanced overall accuracy and F-measure of all classes, but also effectively reduced the number of features and overhead of classification. By comparison with chi-squared in [18] and information gain in [19], the proposed method has lower computational complexity. Chi-square and CFS only considered the correlation between classes and features, while information gain merely take into consideration the amount of information each feature brought. Nevertheless, our method takes into account the coefficient of variation with internal and external class perspective. Theoretically, the features selected by the proposed approach are more effective. Relatively, the method based on coefficient of variation is more efficient, reasonable and robust.

V. CONCLUSION

In this paper, we proposed a new feature selection approach, which is based on coefficient of variation. Moreover, the fine-grained classification of web browsing was first proposed. It has been demonstrated that the presented method has lower computational complexity than chi-squared and information gain. Experimental results showed that the proposed method can achieve higher accuracy of classification than existing methods. Therefore, our approach based on coefficient of variation is an effective and reasonable feature selection method.

In future work, the redundancies of the proposed algorithm can be further reduced by taking account of the correlation of features. We will also try to evaluate the proposed algorithm with some publically available data sets.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (No.61271233), and the HIRP program of Huawei Technology Co. Ltd.

REFERENCES

- [1] Dainotti, Alberto, Antonio Pescapè, and Kimberly C. Claffy. "Issues and future directions in traffic classification." *Network, IEEE* 26.1 (2012): 35-40.
- [2] Joao V. Gomes, et al, Identification of peer-to-peer VoIP sessions using entropy and codec properties, *IEEE Transactions On Parallel And Distributed Systems*, Vol. 24, No. 10, October 2013: 2004-2014.
- [3] Patwary, Md. Mostofa Ali, et al, PANDA: Extreme scale parallel k-nearest neighbor on distributed architectures, *Proceedings - 2016 IEEE 30th International Parallel and Distributed Processing Symposium, IPDPS 2016*, p 494-503, July 18, 2016.
- [4] Quoc, Do Le, et al, Scalable network traffic classification using distributed support vector machines, *Proceedings - 2015 IEEE 8th International Conference on Cloud Computing, CLOUD 2015*, p.1008-1012, August 19, 2015.
- [5] Singh, Hardeep, Performance analysis of unsupervised machine learning techniques for network traffic classification, *International Conference on Advanced Computing and Communication Technologies, ACCT*, v 2015-April, p 401-404, April 2, 2015.
- [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] Crotti M, Dusi M, Gringoli F, et al (2007) Detecting http tunnels with statistical mechanisms. In: *Communications International Conference on IEEE*, pp. 6162-6168.
- [9] Augustin B, Mellouk A (2011) On traffic patterns of HTTP applications. In: *Global Telecommunications Conference (GLOBECOM 2011)*, pp. 1-6.
- [10] Bissias G D, Liberatore M, Jensen D, et al (2005) Privacy Vulnerabilities in Encrypted HTTP Streams. In *Proceedings of Privacy Enhancing Technologies Workshop* 3856:1-11.
- [11] Archibald R, Liu Y, Corbett C, et al (2011) Disambiguating HTTP: classifying web applications. In: *Wireless Communications and Mobile Computing Conference (IWCMC)*, pp. 1808-1813.
- [12] Kaoprakon S, Visoottiviset V (2009) Classification of audio and video traffic over HTTP protocol. In: *Communications and Information Technology, 9th International Symposium on IEEE*, pp. 1534-1539.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] Julie A. Jacko, Andrew Sears, Michael S. Borella (2000) The effect of network delay and media on user perceptions of web resources. *Behaviour & Information Technology* 19(6):427-439.
- [17] Yu J, Lee H, Im Y, et al (2010) Real-time classification of Internet application traffic using a hierarchical multi-class SVM. *KSI Transactions on Internet and Information Systems (TIIS)* 4(5): 859-876.
- [18] Lei D, Xiaochun Y, Jun X (2008) Optimizing traffic classification using hybrid feature selection. In: *Web-Age Information Management, the Ninth International Conference on IEEE*, pp. 520-525.
- [19] Azhagusundari B, Thanamani A S (2013) Feature selection based on information gain. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* 2013: 2278-3075.
- [20] Piraisoody G, Huang C, Nandy B, et al (2013) Classification of applications in HTTP tunnels. In: *Cloud Networking (CloudNet)*, pp. 67-74.



Yuning Dong (M'07), received his B.E and M.E degrees from Nanjing University of Posts & Telecommunications (NUPT), Ph.D degree from Southeast University, all in Electrical Engineering, and M.Phil degree in Computer Science from The Queen's University of Belfast (QUB). He is currently professor with College of Communications and Information Engineering, NUPT. He was a British Council postdoctoral fellow at Imperial College London, 1992-93; a visiting scientist at University of Texas, 1993-95; a research fellow at QUB and University of Birmingham, 1995-98. His research interests include wireless networking, multimedia communications and network traffic identification.



Kai Wang, born in 1990, received his M.E degree in signal and information processing from the Nanjing University of Posts and Telecommunications. His main research interests are network traffic identification and multimedia communications.