Unit traffic classification and analysis on P2P video delivery using machine learning

Rina Ooka, Takumi Miyoshi, and Taku Yamazaki
Graduate School of Engineering and Science, Shibaura Institute of Technology,
307 Fukasaku, Minuma-ku, Saitama-shi, Saitama 337–8570, Japan
a) mf19018@shibaura-it.ac.jp

Abstract: This paper proposes a new traffic classification method for peer-to-peer-based video streaming (P2PTV) by using machine learning. Since P2PTV is a promising technology for video delivery, it is important to comprehend its traffic characteristics. Our previous studies classified and analyzed traffic obtained in viewing P2PTV on a per content basis. However, users’ participation would be dynamic, and then the traffic characteristics may change momentarily. In the proposed method, we divide per-content traffic into short-time data units and then classify them by machine learning. From the experimental results on 80 contents data, six types of unit traffic patterns appear. Moreover, we also analyze their occurrence patterns.

Keywords: P2PTV, machine learning, traffic analysis, clustering

Classification: Network

References
[1] N. Takeuchi, T. Miyoshi, and O. Fourmaux, “Characteristic analysis and modeling of P2PTV traffic,” Proc. 3rd Int. Conf. Inform. Sci. & Security (ICISS2016), Pattaya, Thailand, 4 pages, Dec. 2016. DOI:10.1109/ICISSEC.2016.7885846
[2] K. Hayashi, R. Ooka, T. Miyoshi, and T. Yamazaki, “P2PTV traffic classification and its characteristic analysis using machine learning,” Proc. 20th Asia-Pacific Netw. Operations & Management Symp. (APNOMS2019), Matsue, Japan, 6 pages, Sept. 2019.
[3] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” Proc. 5th Berkeley Symp. Math. Statist. & Prob., vol. 1, pp. 281–297, 1967.
[4] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” J. Comput. Appl. Math., vol. 20, pp. 53–65, Nov. 1987. DOI:10.1016/0377-0427(87)90125-7
[5] PPStream. [Online]. Available: http://www.pps.tv/
1 Introduction

Peer-to-peer-based video delivery service (P2PTV) has attracted attention with an increase of users who enjoy video distribution services. Although P2PTV can distribute the data delivery load concentrated to the video servers, unexpected traffic may be generated between Internet service providers (ISPs) since P2P mechanism does not consider the physical network topology nor peers’ locations.

From the viewpoint of network management, it is necessary to understand the characteristics of P2PTV traffic in advance. In our previous studies, we have obtained P2PTV traffic in watching video contents for a long time and then analyzed the characteristics by classifying the obtained traffic data [1, 2]. However, users would participate in or leave from P2PTV services dynamically, and then the traffic characteristics may change from moment to moment. The traffic classification and analysis on a per content basis must be therefore insufficient.

In this paper, we propose a P2PTV traffic classification method. The proposed method divides traffic generated by P2PTV application into short-time data pieces to create time series data, and then classifies these data by machine learning. We also analyze the unit traffic patterns from the classification results of 80 P2PTV traffic data.

2 Unit traffic classification method

This section explains a unit traffic classification method to analyze traffic occurrence patterns. The purpose of this paper is to obtain the knowledge of P2PTV traffic characteristics and applications behavior through the statistical traffic classifications on a user peer. The analysis even on one peer is enough useful for network providers since it would give suggestions to estimate which directions the traffic goes to or comes from. In the method, each per-content P2PTV traffic is divided into short-time data pieces and then classified by machine learning. For a traffic classification, we implement $k$-means clustering method [3] and the mean silhouette value [4].

2.1 Classification procedure

We describe the classification procedure of P2PTV traffic in the proposed method. First, as a preprocessing, P2PTV traffic data of each content viewed for a long time is divided into small data pieces per unit time (called data units hereafter). When the number of data units created from the original traffic data $i$ is denoted by $n_i$, $n_i = \lceil t_i/T_D \rceil$, where $t_i$ represents the data length (viewing time) of content $i$, and $T_D$ represents the unit time for dividing the traffic data. $\lceil x \rceil$ is the ceiling function that represents the smallest integer greater than or equal to $x$.

Next, we create the training data set from the feature quantities, which are extracted from the data units. The feature quantities are described in 2.2. Then, we classify the training data created above by using $k$-means clustering method. When we use $k$-means method, however, we have to determine the appropriate value of $k$ in advance. To find the optimal $k$, we use the mean silhouette value to examine the clusters. The silhouette value of training data $i$, denoted by $s_i$, is defined as follows:
where $a_i$ represents the degree of cohesion that indicates the average distance to other training data in the same cluster, and $b_i$ represents the degree of separation that indicates the minimum average distance to training data in other clusters. The mean silhouette value is calculated as the average of $s_i$ for all training data. The silhouette value is in the range $[-1, 1]$. Larger mean silhouette value represents better clustering; i.e., higher inter-cluster cohesion and intra-cluster separation. In this paper, we change $k$ in $k$-means method from 1 to 30 and search the optimal value of $k$ when the mean silhouette value becomes the largest.

### 2.2 Feature quantities extraction

For clustering P2PTV traffic by machine learning, it is necessary to quantify the features. In our previous study, 18 types of feature quantities are extracted from P2PTV traffic data [2]. However, some of them are difficult to extract in a short time period. Therefore, we use only the useful feature quantities that can be extracted even in a short time. Consequently, we determine 6 types of feature quantities: the average number of simultaneously connected peers, the maximum number of simultaneously connected peers, the total number of connected peers, the average throughput, the maximum throughput, and the coefficient of variation of throughput (defined as the standard deviation divided by the average). Note that the number of simultaneously connected peers and the throughput are measured every second. In addition, different from the previous study, we focus on both received and sent data traffic to realize more efficient classification. Thus, 12 types of feature quantities are extracted.

### 3 Evaluations

#### 3.1 Experimental setup

This paper focuses on PPStream [5], one of popular P2PTV applications, and especially on on-demand video delivery. As the traffic data set, considering the computational time, we randomly selected 80 video traffic data from 390 data in [1], which had been captured by a packet sniffer application, Wireshark, in watching PPStream video contents on the measurement PC just as a user peer. The unit time for dividing traffic data $T_D$ was set to 30 or 10 seconds. Although only about one-fifth of the original data set were utilized, the total numbers of data units created from per-content traffic were considerable: 5,080 ($T_D = 30$) and 14,476 ($T_D = 10$). Note that only the packets of not less than 1,000 bytes were used in this experiment to focus on video data by eliminating small-size control packets.

#### 3.2 Classification results

We obtained as results of clustering that the numbers of clusters are two to six when $T_D = 30$, and two to seven when $T_D = 10$. Comparing the clustering results of 80 video traffic data, most of the training data were classified into two or three clusters regardless of $T_D$; the occupancy rates were about 83% and 90% of the total video...
contents when $T_D = 30$ and $T_D = 10$, respectively. From these characteristics, we consider that most of the traffic can be expressed as a combination of two or three unit traffic patterns in PPStream. In the case of video contents classified into four to seven clusters, there existed two or more clusters for stable traffic according to the differences in throughput. On the other hand, bursty traffic was always classified into the same cluster.

To analyze the characteristics of traffic data included in each cluster, Fig. 1 shows six representative examples of traffic sent and received at the measurement PC in $T_D = 10$ as the relationship among elapsed time, the number of simultaneously connected peers, and the throughput. The characteristics of P2PTV unit traffic are confirmed as: (a) bursty traffic of received data, (b) stable traffic of received data, (c) equivalent throughput in both sent and received data, (d) bursty traffic of sent data, (e) stable traffic of sent data, and (f) no traffic generated.

### 3.3 Analysis of unit traffic patterns

We examined unit traffic patterns included in each cluster. Fig. 2(a) shows the whole traffic data when viewing a video content, and Fig. 2(b) and (c) show the numbers of training data as well as the representative traffic graphs classified into three and four clusters when $T_D = 30$ and 10, respectively. $C_1$ is the largest cluster of the very low-throughput stable sent traffic with almost no received data. On the other hand, $C_2$, $C_3$, and $C_4$ are the clusters of the received traffic feature; $C_2$ is low-throughput ($<500$ Kbytes/sec) stable received traffic with little sent data, $C_3$ is bursty traffic without sent data, and $C_4$ is high-throughput ($\geq 500$ Kbytes/sec) stable received traffic with little sent data. When $T_D = 30$, all the stable received traffic with little sent data was classified into $C_2$ regardless of received throughput.

To examine how data units classified into each cluster appear, Fig. 3 shows the color-coded transition of clusters at each unit time $T_D$ for the same video traffic shown in Fig. 2. Here we qualitatively evaluate the result of $T_D = 10$, which indicates finer cluster transition than $T_D = 30$. The figure shows that $C_4$ appears first. This represents that video data should quickly come from the sufficient

---

**Fig. 1.** Representative traffic features when $T_D = 10$.

(a) Bursty received. (b) Stable received. (c) Stable sent and received.

(d) Bursty sent. (e) Stable. (f) No traffic.
number of neighbor peers at the very beginning. Next, the phase shifts to $C_2$. We consider that large subsequent video data arrive continuously and are cached from fewer peers. $C_3$ seems to be observed at the edge of the $C_2$ phase. After the data

Fig. 2. Representative unit traffic included in each cluster.

Fig. 3. Unit traffic occurrence patterns.
cache is enough filled, $C_1$ follows, where video data are scarcely received. Thus, the unit traffic occurrence patterns can accurately represent the video traffic shown in Fig. 2(a). Moreover, the unit traffic treated as one data unit when $T_D = 30$ can be divided more finely into smaller data units when $T_D = 10$. The bursty unit traffic, in particular, can be divided into a bursty part or a non-bursty part by smaller $T_D$. Thus, we consider that the traffic classification with smaller time intervals realizes better and detailed analysis to understand P2PTV traffic in the proposed method.

4 Conclusions
This paper proposed a classification method for P2PTV traffic using clustering. The obtained results showed that there exist six representative unit traffic patterns, most of the data units can be classified into two clusters, and more detailed analysis on unit traffic occurrence patterns will be possible when the unit time $T_D$ is smaller.

As a future work, we will try to improve the classification accuracy by changing the feature quantities and analyze the unit traffic occurrence patterns generated by other P2PTV applications. Moreover, we will consider an extended method to classify the whole P2PTV traffic using time-series unit traffic patterns.

Acknowledgments
This study was partly supported by JSPS KAKENHI Grant Number 17K06441.