Analysis and Practicality of Traffic Flow Classification Method using Binned Time-Series Data

Shuichi Nawata, Hideyuki Koto, Takeshi Kitahara, Shigehiro Ano
KDDI R&D Laboratories Inc.
2-1-15 Ohara, Fujimino-shi,
Saitama, 356-8502 Japan
{nawata, koto, kitahara, ano}@kddilabs.jp

***Abstract*— It is important for network operators to know how users use the network to plan and design network facilities. In this paper, we propose flow classification method using binned time-series data in consideration of practicality. In this paper, we evaluate the proposed method by using mobile data traffic and show the method can maintain classification accuracy as compared with method using the whole flow information. We also confirm applicability to classification of encrypted flow.**

***Keywords*— flow classification; binned data; encrypted communication;**

I. INTRODUCTION

Network traffic has grown yearly, and [1] predicts that global mobile data traffic will increase 10-fold from 2014 to 2019. Network operators need to understand how users use the network because they must strengthen their network facilities to maintain communication quality. Traffic classification technique is one of the ways to know how the network is used. In recent years, a number of researchers have studied about ways to classify traffic data into types of applications. These studies include a clustering technique based on swarm intelligence to classify accurately with small computational complexity [2], the classification technique with LPC (Linear Predictive Coding) cepstrum to reduce influence of capture time and environment [3], and the method of application identification based on characteristic change by encryption of traffic to classify the mixed traffic data of encrypted/normal communications [4].

Meanwhile, in practical environments, Network operators have to analyze enormous number of flows. Therefore, such as simple implementation, small computational complexity and memory consumption during monitoring is an important issue. However, since the diversity of application usage by the users and devices are large, it's difficult to estimate the duration of every flow beforehand. Accordingly, in this paper, we propose a flow classification method using binned time-series data. For implementation and computational complexity, the proposed method monitors a fixed time length from the beginning of a flow. In addition, we evaluate the proposed method mobile data traffic of smartphones. Furthermore, since we can perform frequency analysis using binned time-series data, we also evaluate the method using frequency properties that show periodicity and density of packets as features of flow.

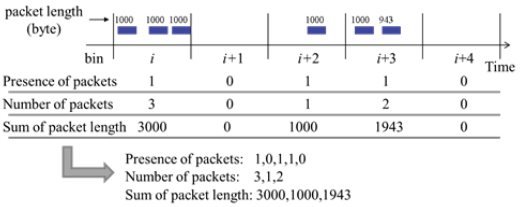

Fig. 1. Example of count of flow information in each bin

II. PROPOSED METHOD

In this section, we explain the process flow of the proposed method which consists of three steps: data format conversion, feature extraction, and classification.

A. Data format conversion

First, we explain preprocessing of flow classification.

1) Flow identification: We define a flow as a series of packets that have the same 5-tuple information: source IP address, destination IP address, source port number, destination port number, and protocol obtained from IP header and TCP header. Our method categorizes packets 5-tuple.

2) Partition of flow: After flow identification, we perform bin sepalation of a flow. Fig. 1 shows an example of this step. Different packets are recorded as a same bin where their monitored timings in terms of bin number are the same. The begining of the first bin is set as the begining of a flow. However, we do not count/record packets where their TCP payload size is 0 to omit SYNs, FINs, etc. In other words, we only count data packets to improve classification accuracy [2]. Next, we count and record several features of packets in a same bin: presence of transmitted/received packets, the number of transmitted/received packets, and the sum of transmitted/received packet lengths in each bin for each flow. Finally, we limit the maximum number of bins, i.e. fixed time length, as M.

B. Feature Extraction

1) Feature calculation: We calculate features of flow from binned time-series data to use them in flow classification. Table I shows a list of the calculate features. We devise two sets of features. One is the statistical information of the number of packets and sum of packet lengths. In addition, we utilize frequency properties of the observed data. As an

TABLE I. LIST OF FEATURES FOR BINNED TIME-SERIES DATA

Attribute	Feature name	Description
(1) Statistics calculated from binned time-series data (18 features)	(sum/avg/std/cv)_pkt_(up/down)	sum/average/standard deviation/ coefficient of variation of number of transmitted/received packets per bin
	ratio_pkt	natural logarithm of total number of received packets divided by one of transmitted packets per bin
	(sum/avg/std/cv)_sz_(up/down)	sum/average/standard deviation/ coefficient of variation of sum of transmitted/received packet lengths per bin
	ratio_sz	natural logarithm of total size of received packets divided by one of transmitted packets per bin
(2) Power spectrum calculated from binned time-series data (12**N* features)	prsnc_(up/down)_f*N*	frequency of the *N*th largest power based on the presence of transmitted/received packets
	prsnc_(up/down)_p*N*	the *N*th based on the presence of transmitted/received packets
	pkt_(up/down)_f*N*	frequency of the *N*th largest power based on the number of transmitted/received packets
	pkt_(up/down)_p*N*	the *N*th based on the number of transmitted/received packets
	sum_sz_(up/down)_f*N*	frequency of the *N*th largest power based on the sum of transmitted/received packet length
	sum_sz_(up/down)_p*N*	the *N*th based on the sum of transmitted/received packet length

example in this paper, we perform power spectrum analysis where periodicity and density of transmitted/received packets could be analyzed as features of flow.

C. Classification

Finally, we classify flows into application types using the features of Table I. We select features of flows and apply cross-validation as bellow. These processing are performed on all monitored flows. In this paper, we use Weka, which is open source software, to apply algorithms of machine learning for classification [6].

1) Feature Selection: We select and omit features to improve classification accuracy. We use a rapper approach with the C4.5, which is one of the general decision tree algorithms of machine learning, for an evaluator and use best first search algorithm for a search.

2) Cross Validation: By 10-fold cross-validation, we inspect classification accuracy. We also use C4.5 as classification algorithm. This is only an example and any other suitable machine learning method could be used.

III. EVALUATION AND DISCUSSION

In this section, we evaluate classification accuracy of the proposed method. To evaluate the method, we use four types of popular applications: audio streaming, file download, video streaming, and web browsing. We analyze mobile data traffic of smartphones with 397 flows for each application and evaluate classification precision of 1,588 flows in total. As a comparison to the proposed method, we also calculate the

TABLE II. LIST OF FEATURES FOR ALL PACKETS FLOW

Attribute	Feature name	Description
Flow information (5 features)	tcp_duration	duration of TCP session
	tcp_num_pkt_up/down	number of transmitted/received packets of TCP session
	tcp_len_up/down	sum of transmitted/received packet length of TCP session

TABLE III. RESULTS OF CLASSIFICATION
(1) USING THE FEATURES ABOUT FLOW LEVEL INFO

Class	F-measure	TP Rate	FP Rate	Precision	Recall
AUDIO	0.834	0.841	0.059	0.827	0.841
File DL	0.868	0.844	0.034	0.893	0.844
VIDEO	0.941	0.950	0.023	0.933	0.950
WEB	0.920	0.929	0.030	0.911	0.929

(2) PROPOSED METHOD (*M*=1000)

Class	F-measure	TP Rate	FP Rate	Precision	Recall
AUDIO	0.846	0.836	0.047	0.856	0.836
File DL	0.884	0.884	0.039	0.884	0.884
VIDEO	0.965	0.970	0.013	0.960	0.970
WEB	0.921	0.927	0.029	0.915	0.927

classification accuracy where the statistical features, in Table II, for the wide time length of flows, i.e. not limited bins. We use F-measure as a performance metric to judge precision and recall in a comprehensive manner. F-measure F is harmonic average of precision P and recall R and is defined in following expressions.

$$F = 2 \cdot P \cdot R/(P + R)$$
$$P = True\ Positives/(True\ Positives + False\ Positives)$$
$$R = True\ Positives/(True\ Positives + False\ Negatives)$$

A. Classification using statistics of binned time-series data

First, we check whether the proposed method maintain the same classification accuracy as the case where the whole time length of flow information is used. We set the time length of a single bin as 10 msec and M, the limit of maximum bins, as 1,000 which is 10 seconds. Table III shows classification accuracy for the two methods. Comparing the F-measure, the values of the proposed method is higher (0.001~0.024) than the method using the whole flow information.

Next, we inspect the effect of M to classification accuracy of the binned data. We set M as 1,000 and 6,000 (i.e. 10 and 60 seconds) and compare their classification accuracy. We show the results of classification in Fig. 2. As the figure shows, both accuracies are approximately equal. This means that, for the evaluated data sets, increasing the time length of bins, i.e. the observed data, does not improve classification accuracy. To further evaluate the reason for this, we analyzed the flow duration of the data sets used. Fig. 3 shows CDF (cumulative distribution function) of while flow duration of each application. As the results show, for example, although the CDF for audio streaming increases from approx. 77 % to 90 % when durations are 10 to 60 seconds, the corresponding F-measures does not improve or ever decrease slightly. This shows that just increasing M is not effective to improve accuracy, thus limiting M could save resources while maintaining accuracy.

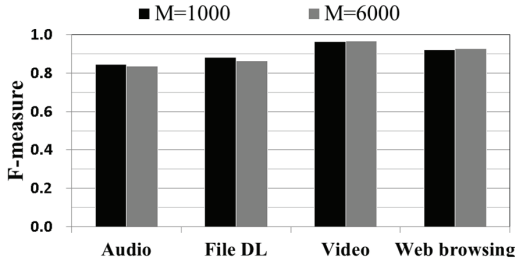

Fig. 2. Results of classifications using statistics calculated for M=1,000 and M=6,000

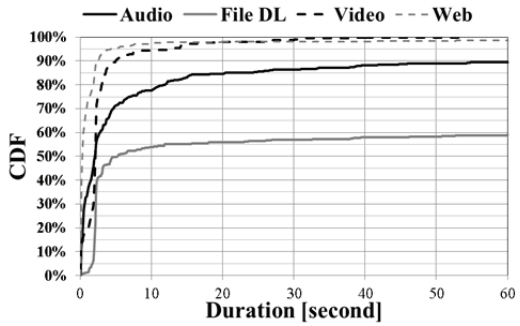

Fig. 3. CDF of while flow duration

B. *Effect of power spectrum*

In this section, we show an example where frequency analysis is utilized. In this paper, we use power spectrum which is one of the frequency analysis in one instance. Fig. 4 shows the F-measures as a result of traffic classification using the power spectrum and the whole flow information, respectively. Note that N indicates the number of Nth largest power and corresponding frequency. M is set at 1,000 (10seconds). As the figure shows, in the case of using power spectrum as features, the accuracy for N=10 is the highest. In addition, comparing all the results, they are more or less similar. The F-measure, however, for video streaming improved by 2.9 % for N=10 compared to that of flow information. This suggests that frequency analysis method could improve accuracy of specific applications where the frequencies of traffic patterns are distinctive. Note that we compared the accuracy for M=6,000 and confirmed that classification accuracy did not have a big difference.

C. *Application for encrypted communications*

Finally, we investigate the applicability of the proposed method to encrypted flows. For this evaluation, we changed the data sets for video streaming with encrypted data, i.e. HTTPS flows. We compare the classification accuracy of the flow information with that of the statistics calculated from binned time-series data. Table IV (1) and (2) show the results are similar. Both accuracies are also at the same level as those of Table III. These results indicate that our proposed method could identify flows of video streaming with high accuracy regardless of encryption.

IV. CONCLUSION

In this paper, we proposed a flow classification method using binned time-series data. The proposed method monitored a fixed time length from the beginning of a flow in consideration of implementation and computational complexity. We evaluated the proposed method using mobile data traffic of smartphones. Our results showed that the classification accuracy of proposed method was at the same level of that of method using flow information of the whole time length. In addition, we confirmed the applicability to encrypted flow classification for specific application.

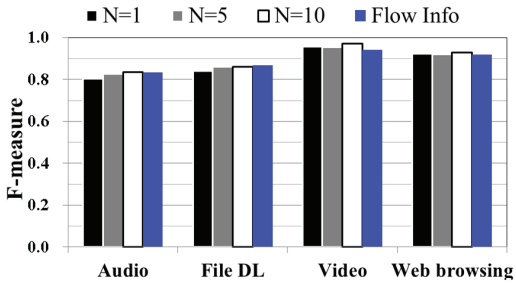

Fig. 4. Results of classifications using power spectrum analysis and using flow information

TABLE IV. RESULTS OF CLASSIFICATION
(1) USING THE FEATURES ABOUT FLOW

Class	F-measure	TP Rate	FP Rate	Precision	Recall
AUDIO	0.830	0.836	0.06	0.824	0.836
File DL	0.870	0.846	0.033	0.896	0.846
VIDEO (encrypted)	0.939	0.952	0.025	0.926	0.952
WEB	0.921	0.927	0.029	0.915	0.927

(2) USING STATISTICS ABOUT THE TIME-SERIES DATA BASED ON BIN FEATURE EXTRACTION (M=1,000)

Class	F-measure	TP Rate	FP Rate	Precision	Recall
AUDIO	0.832	0.811	0.046	0.854	0.811
File DL	0.881	0.882	0.040	0.879	0.882
VIDEO (encrypted)	0.966	0.970	0.013	0.963	0.970
WEB	0.941	0.960	0.027	0.923	0.960

REFERENCES

[1] Cisco Systems, Inc., "Cisco Visual Networking Index," Feb. 2015. [Online] Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html

[2] T Sue, Y. Ohsita, and M. Murata, "Application and Evaluation of Clustering Methods using Swarm Intelligence to Traffic Classification," Technical report on IEICE, vol. 114, no. 29, IN2014-15, pp. 37-42, May 2014. (In Japanese)

[3] M. Ichino, H. Maeda, N. Komatsu, K. Takeshita, M. Tsujino, H. Hasegawa, and H. Yoshino, "The Internet Traffic Classification Method Using LPC Cepstrum," IEICE transactions on communications B, vol. J95-B, no. 7, pp. 835-847, 2012. (In Japanese)

[4] Y. Okada, S. Ata, N. Nakamura, Y. Nakahira, and I. Oka, "Application Identification from Encrypted Traffic based on Characteristic Change by Encryption," in Proceedings of the IEEE International Communications Quality and Reliability Workshop (CQR 2011), (Naples, Florida, USA), May 2011.

[5] S. Nawata, H. Koto, N. Fukumoto, and H. Yokota, "A Study of Method for Extracting Feature of Internet Traffic by Using Power Spectrum," in Proceedings of 2014 IEICE General Conference, B-6-76, Mar. 2014. (In Japanese)

[6] Machine Learning Group at the University of Waikato, "Data Mining Software:Weka," 2014. [Online] Available: http://www.cs.waikato.ac.nz/~ml/weka/index.html