Exploiting Adaptive Packet-Sampling Measurements for Multimedia Traffic Classification

Rongqiang Lin, Ou Li, Qing Li, and Kun Dai
National Digital Switching System Engineering and Technological Research Center, Zhengzhou, China
Email: linrq0208@126.com; {zzliou, liqing0206, daikun_1223}@163.com

Abstract—With the huge amount of ubiquitous multimedia data transmitted in nowadays Internet, the use of packet sampling for traffic measurements has become widely employed for network operators. In this paper, we present an adaptive packet sampling technique from the classification perspective, the main sampling principle of which is to select as many packets with low occurrence rate as possible based on two useful features for multimedia traffic: Packet Size (PS) and Packet Inter Arrival Time (IAT). We build a model of the ideal packet sampling technique for classifying multimedia traffic, which adjusts adaptively the sampling probability of selecting packets according to PS and IAT predicted simultaneously by multi-output support vector regression, and define general indexes for evaluating the sampling performance of the proposed approach. We compare our approach with other sampling methods and evaluate their impact on the performance of traffic classification using two machine learning methods with real multimedia traffic data. The experimental results show that this approach has good sampling performance and is able to enhance the performance of the traffic classification methods.

Index Terms—Sampling, traffic classification, multimedia

I. Introduction

The Internet has proven to have an extraordinary capability of adapting to new services, transferring from the initial pure datagram paradigm which makes no guarantees of the delivery of each message to the real multi-service infrastructure [1]. The evolution of the Internet has induced various multimedia applications (such as Skype, MSN, IPTV, etc.) [2], [3], and the mobility of users will likely translate into the mobility of services. The explosion of the multimedia services implies that we should reconsider the profound connotation of the data traffic. Moreover, the widespread usage of application layer protocols directly translates into a much higher variability of the data traffic injected into the Internet.

As a side effect of the rush of the multimedia services and the increasing transmission speed, companies and Internet Service Providers (ISPs) must deal with ever-growing traffic and the huge amount of measurements, and how to collect, store and process the consequent traffic data is one of the most serious challenges we face today. Therefore, in order to reduce the amount of data to a manageable size, packet sampling has become compulsive for effective passive network measurements especially in the core of the network. Naturally sampling inherently implies some loss of information and several studies have focused on the impact of different sampling policies on traffic measurement [4]-[7]. However, to our knowledge, only a few works in the literature discuss the effect of sampling on the performance of various networking activities, such as monitoring, service-level agreement compliance, anomaly detection and traffic classification [8]-[15].

Among the different applications of traffic measurement, traffic classification has recently attracted considerable attention. Traditional traffic classification based on well-known transport layer port numbers becomes unreliable, due to the fact that emerging Internet applications tend to mask their identifications by using random ports [3]. There exist three other classification methods in recent years: (1) DPI (deep packet inspection) which searches for the known signatures in the packet payload; (2) host behavior based methods to dig the hidden connection patterns between hosts; (3) machine learning based methods using statistical features of traffic such as packet size, flow duration [16]. Although DPI-based schemes can provide high performance, they become null when dealing with encrypted applications and may be involved in the government regulations. The effectiveness of host behavior based approaches depends on topological locations and traffic mixes, but they are invalid in identifying the application type for single packets [17]. Recent research on traffic classification tries to identify network applications by learning statistical patterns in externally observable features of packets or flows [2]. Such statistical approaches based on machine learning have thus been considered more promising and robust to encryption, privacy, protocol obfuscation [16]-[18].

However, as sampled data are becoming the only kind of data available for multimedia traffic, the question is whether classification is still possible with such reduced information from sampled data. Still, several studies have focused on the impact of traffic sampling on traffic classification performance [14], [15], which have shown

Manuscript received September 10, 2014; revised December 30, 2014.

This work was supported by National Natural Science Foundation of China (NO. 60872043) and National 863 plans project (NO. 2009AA01Z207).

Corresponding author email: linrq0208@126.com.

doi:10.12720/jcm.9.12.971-979

that the classification performance degrades drastically due to sampling. So, it is important to select these packets carrying the most useful information in order to achieve the least impact of traffic sampling on multimedia traffic classification.

The main contributions of this paper as following:

- We present a novel packet sampling method from the classification perspective. (1) This approach adjusts adaptively the sampling probability according to PS and IAT predicted simultaneously by multi-output support vector regression. (2) The sampling policy could sample the representative packets of the multimedia traffic according to these two features, while ensuring the classification performance and reducing the amount of samples to a manageable size.
- We define general evaluating indexes and evaluate the extent of the degradation of the information content conveyed by PS and IAT. We compare our approach with random sampling and reveal their impact using two machine learning classification methods with real multimedia traffic. Our analysis and experimental evaluation show that the proposed method can also be appropriate to refine the input data by reducing highly superfluous packets.

This presented packet sampling method provides the following benefits:

- It selects as many packets with low occurrence rate as possible and throws theses packets with high occurrence rate away with the same values of $\langle \mathrm{PS}, \mathrm{IAT} \rangle$, thus gets the representative samples for training the classifiers.
- Taking into account the behavioral characteristics of the traffic by adjusting the sampling probability according to its historical information, this approach can be applied on high speed network. When dealing with high-speed traffic, traditional sampling methods, such as systematic or random method, cannot help but to discard large numbers of useful packets because their sampling granularity couldn't trace the arrival rate of the real traffic.

We characterize the distortion introduced by sampling by means of two statistical metrics, namely the sampling loss and the sampling loss coefficient between the sampled and unsampled values. As for the impact of sampling on traffic classification, we evaluate the classification performance through two classifiers: support vector machine (SVM), a widely known supervised classification algorithm, and VOVClassifier, an automated self-learning system that classifies traffic data by extracting features from frequency domain [3].

This remainder of this paper is organized as follows. First an overview of the related work is found in Section II. Then, we describe the methodology used to process these data in Section III, in particular, the ideal packet sampling model, Multi-output support vector regression, the sampling policies and the classification algorithm. Section IV analyzes the sampling performance and Section V presents the impact of sampling on multimedia traffic classification. We conclude our paper and discuss the future work in Section VI.

II. Related Work

The need for sampling for network measurements was already identified more than a decade ago in [19], [20]. Duffield [21] offers an exhaustive overview of uses of sampling in this field. In the following we overview the most important pieces of work on this topic, both on sampling itself and on its effect on traffic classification.

Researchers have categorized sampling methods into two classes: packet-sampling and flow-sampling. Packet-sampling methods work on the levels of network packets [10], and each packet is selected according to the selection scheme, which can be systematic or random. We can differentiate sampling techniques according to the selection trigger, based on the amount of time or number of packets between two different sampling events. Claffy *et al.* [6] showed that time-based triggers are less robust than packet-based ones, because they suffer from the bursty nature of network traffic. Both the Internet IETF (Internet Engineering Task Force) working groups, IPFIX (IP Flow Information Export) [22] and PSAMP (Packet Sampling) [23], have recommended the use of packet sampling. Compared with traffic sampling, the main advantage of packet sampling is the decreased requirements for memory consumption and CPU power on routers as well as the possibility to monitor higher network speed. As for flow sampling, the objective traffic is aggregated into network flows and the sampling itself is applied not to the particular packets, but to the whole flows. The main benefit of flow sampling methods is that they have better performance compared to packet sampling [24], but they require more memory and CPU power. Smart sampling [7] and sample-and-hold [25] methods were then introduced to reduce the memory requirements. Both of these methods are focused on accurate traffic estimation for large flows. In this paper, we emphasize the packet sampling method.

Although packet sampling is easier to implement, it introduced some challenging problems in flow statistics [7], [24]. First, the packet rate of a flow (the packets/flow) changes a lot during the duration of this flow, which makes it difficult to define mice flows and elephant flows. Second, the arrival time of a flow is dynamic. One of the main reasons for this is the transmission delay due to router queue congestion and the time-varying channels of the transmission paths, etc. Finally, the duration of a flow will vary a lot over time and it stay active for a random duration. So it is important to adapt the sampling probability according to changes in the traffic. Some works have proposed to make the sampling probability adaptive [5], [8], [26], for instance, to the traffic load, to reduce the estimation error of some traffic metrics. However these works have rarely considered the impact of packet sampling and the performance of traffic classification techniques together. This motivates us to develop an adaptive packet sampling technique for traffic classification, especially for the widely adopted machine learning methods which make use of statistical information of a flow. Bartos *et al*. [10] proposed an adaptive flow sampling technique for anomaly detection, which defined sampling probability through primary features and secondary features. Different from the above work, our main focus is on packet sampling technique for traffic classification based on different sampling principles.

Some recent papers do not focus only on sampling itself, but also on their effect on traffic classification. Erman *et al*. [27] analyze how sampling methodology influences the selection of both elephant flow and mice flow in the training dataset, aggravating the traditional class imbalance problem. Park *et al*. [28] investigate the sampling effect on Reduced Error Pruning Tree classifiers. Instead, Jiang *et* al. [13] investigate the sampling effect on a lightweight traffic classification approach using Naïve Bayes on NetFlow records, and varying the sampling rate, and find that packet sampling does not worsen the results (rather, accuracy may increase under heavy sampling) and suggest this may be due to an artifact of packet sampling. Carela-Espanol *et al.* [14], which studies the accuracy of statistical traffic classification based on NetFlow sampled data. Tammaro *et al*. [15] assess the impact of packet sampling on various network monitoring-activities, with a particular focus on traffic characterization and classification. Finally, these works consider only traditional sampling techniques, while in our work we present an adaptive, feature-aware sampling policy which is specially defined for multimedia traffic classification.

III. Methodology

We now describe the methodology followed in the experimental study. First, we describe the model of the ideal packet sampling designed for traffic classification. Second, we detail the multi-output support vector regression used for forecasting PS and IAT. Third, we present the sampling policy we apply to packet-level traces. Finally, we describe the classification algorithm we employed to test the effect of sampling on traffic classification.

A. Ideal Packet Sampling Model

In the first part, we build a model of the ideal packet sampling specially designed for multimedia traffic. Visually, the goal of the ideal packet sampling is to select packets in such a way that information loss evaluated by indexes is minimal. However, different traffic features may have different impact on classification performance when dealing with the ideal packet sampling for traffic classification, which is due to the fact that traffic classification methods use only some of these features derived from packets as input data. Therefore, some features are more important and representative than others for classification purpose. Fan *et al.* [3] demonstrates that multimedia traffic show strong regularities in these two features (PS and IAT), using which the classifier can achieve a high accuracy of the multimedia traffic classification. This implies that these two features are the very important information contents of a flow, and there exist correlation between PA and IAT. So if we select flows according to PS and IAT, it has a good potential to obtain a representative dataset from the raw traffic, while reducing the amount of data to a manageable size. Therefore, in this paper we select PS and IAT for the important features for their representative, which will be shown to be sufficient to achieve high performance for sampling and classification.

We denote each packet from multimedia traffic as P, which can be identified by features: PS and IAT. In the eyes of traffic classification, it is important to select all the representative packets by using sampling methods. On the other hand, the meaning decreases with the growing number of similar packets with the same values of $\langle \mathrm{PS}, \mathrm{IAT} \rangle$ already in the set from the classification viewpoint, as these packets may be of the same application type with good probability. So the main sampling principle of our approach is to sample as many as packets which have different values of $\langle \mathrm{PS}, \mathrm{IAT} \rangle$. For example, assumed that the number of unsampled packets with different values of $\langle \mathrm{PS}, \mathrm{IAT} \rangle$ is 10000, and then the goal of our sampling approach is to obtain the sampled packets whose sum is close to 10000 with different $\langle \mathrm{PS}, \mathrm{IAT} \rangle$. Here, the meaning of different values of $\langle \mathrm{PS}, \mathrm{IAT} \rangle$ between two packets (P_i and P_j) is that they must satisfy such conditions $\mathrm{PS}_i \notin \left[\mathrm{PS}_j - \theta_{\mathrm{PS}}, \mathrm{PS}_j + \theta_{\mathrm{PS}}\right]$ and $\mathrm{IAT}_i \notin \left[\mathrm{IAT}_j - \theta_{\mathrm{IAT}}, \mathrm{IAT}_j + \theta_{\mathrm{IAT}}\right]$, where θ_{PS} and θ_{IAT} are constants for PS and IAT respectively.

Let U be the original finite unsampled set and S the finite sampled set. Furthermore we will denote $N_U\left(P \mid \mathrm{PS}, \mathrm{IAT}\right)$ and $N_S\left(P \mid \mathrm{PS}, \mathrm{IAT}\right)$ as the summations and numbers of packets related to P through different values of the pair $\langle \mathrm{PS}, \mathrm{IAT} \rangle$ in U and S respectively. And we consider $N_U\left(P \mid \mathrm{PS}, \mathrm{IAT}\right)$ and $N_S\left(P \mid \mathrm{PS}, \mathrm{IAT}\right)$ as relevant information which the ideal packet sampling should be able to preserve. In the following, we will define the sampling loss which describe the loss of numbers of packets between U and S.

Definition 1: The sampling loss is described by distance function:

$$d\left(U,S\right) = \left| N_U\left(P \mid \mathrm{PS}, \mathrm{IAT}\right) - N_S\left(P \mid \mathrm{PS}, \mathrm{IAT}\right) \right| \quad (1)$$

If the value of $d\left(U,S\right)$ in (1) is equal to zero, it means that original set U is sampled without loss of information and it is possible to reconstruct all the packets in U from the classification viewpoint and the classification performance reaches the maximum value.

Definition 2: Let $S_1,...,S_k$ be various sets of packets sampled from U. All the packets in U are reversible if and only if:

$$\lim_{k\to\infty}\sum_{i=1}^{k} d\left(U,S_i\right)=0 \quad \forall P\in U \tag{2}$$

Reversibility ensures complete reconstruction of finite unsampled set U using only the sampled set S, which is a key sampling property for traffic classification. In order to quantify the distortion introduced by sampling, we also consider the following index.

Definition 3: The sampling loss coefficient is defined as:

$$\Phi(U,S)=\frac{N_S\left(P\mid \mathrm{PS},\mathrm{IAT}\right)}{N_U\left(P\mid \mathrm{PS},\mathrm{IAT}\right)}\in\left[0,1\right] \tag{3}$$

If the value of $\Phi(U,S)$ in (3) is equal to 1, it means that there is no loss of information between $N_U\left(P\mid \mathrm{PS},\mathrm{IAT}\right)$ and $N_S\left(P\mid \mathrm{PS},\mathrm{IAT}\right)$. The sampling loss coefficient is another form of expression of $d\left(U,S\right)$. The sampling loss tells us how many the sampled packets diverge from the unsampled packets: the smaller the distortion the better; the sampling loss coefficient, instead, tells us whether a linear dependence exists between the number of packets with different values of $\langle \mathrm{PS},\mathrm{IAT}\rangle$ in U and S.

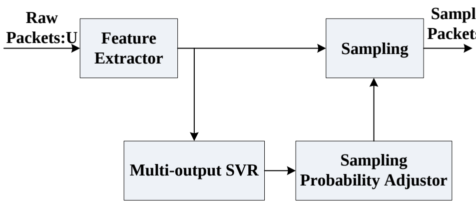

Fig. 1. Architecture of the adaptive packet sampling system

After detailing the idea of ideal sampling, we then describe the overall architecture of our ideal packet sampling system. Fig. 1 shows the system architecture. All packets collected are processed by the Feature Extractor (FE) module, which reorders packets in terms of PS and IAT between packets within any generic flow. The FE output is forwarded to the Multi-output SVR predictor module which predicts PS and IAT simultaneously for the next incoming packet. According to the predicted values of PS and IAT, the sampling probability adjustor module adjusts the sampling probability of the next packet adaptively. Finally, the sampling module gets the sampled set which will be fed to the classifier.

B. Multi-output Support Vector Regression

We are now ready to describe the theory of multi-output support vector regression for traffic prediction. Support Vector Regression (SVR) aims to build a model of the output of a process or system that depends on a set of factors, given input variable $\mathrm{x}\in\mathrm{R}^d$ and output variable $y\in\mathrm{R}$. SVR is traditionally used with only one output, and the multi-output case arises when the output variable is a vector $\mathrm{y}\in\mathrm{R}^d$ instead of a scalar quantity.

The multi-output regression problem could be divided into a number of one-dimensional problems, and in some cases, minimum variance estimation is equivalent to the multi-output regression. However, it is not the case for the prediction of PS and IAT using two one-dimensional SVR for three reasons: (1) There exist relationship between these two output variables as discussed in Subsection A, so it will bring less prediction error using multi-output SVR predicting PS and IAT simultaneously than dividing into two one-dimensional problems. (2) The insensitive zone defined around the estimate will not equally treat every training sample. (3) Predicting PS and IAT respectively using SVR could cause the problem that a sample is a support vector when predicting PS, while the same sample wouldn't be a support vector when predicting IAT, which doesn't make full use of the correlation between PS and IAT.

Therefore, in this paper we introduce Multi-output SVR (MSVR) approach in which a hyper-spherical insensitive zone is defined around the estimate [29]. Given a labeled training data set $\left(\mathrm{x}_i,\mathrm{y}_i\right), i=1,...,n$, where $\mathrm{x}_i\in\mathrm{R}^d$, $\mathrm{y}_i\in\mathrm{R}^k$ and a nonlinear transformation to a higher dimensional space $\phi(\cdot)$, the k-dimensional regression model is $f=\mathbf{W}\phi(x_i)+\mathbf{b}$, where $f=\left[f_1,...,f_k\right]$, $\mathbf{W}=\left[w^{(1)},...,w^{(k)}\right]^T$ and $\mathbf{b}=\left[b^{(1)},...,b^{(k)}\right]^T$ which define the regression model in the high dimensional feature space. According to SVM theory, the MSVR solves the following constrained optimization problem [29]:

$$\begin{aligned} &\min_{\mathbf{w}^j,b^j,\xi_i} \quad \sum_{j=1}^{k}\left\|w^{(j)}\right\|^2+C\sum_{i=1}^{n}\xi_i \\ &s.t. \quad \left\|\mathrm{y}_i-\mathbf{W}\cdot\phi\left(\mathbf{x}_i\right)-\mathbf{b}\right\|^2\le\varepsilon+\xi_i \\ &\qquad \xi_i\ge 0 \qquad i=1,\cdots,n \end{aligned} \tag{4}$$

where $\mathrm{y}_i=\left[y_{i1},...,y_{ik}\right]^T$, C and ε are constants. In order to solve (4), define a Lagrangian:

$$\mathrm{L}=\sum_{j=1}^{k}\left\|w^{(j)}\right\|^2+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\mu_i\xi_i-\sum_{i=1}^{n}\alpha_i(\varepsilon+\xi_i-\left\|\mathrm{y}_i-\mathbf{W}\cdot\phi\left(\mathbf{x}_i\right)-\mathbf{b}\right\|^2) \tag{5}$$

Using the Representer theorem [30], the best solution of (4) can be expressed as a linear combination of the training samples in the feature space, so we get $w^{(j)}=\sum_{i=1}^{n}\beta_i^{(j)}\phi(x_i)=\boldsymbol{\varphi}^T\beta^{(j)}$, where $\boldsymbol{\varphi}=\left[\phi(x_1),...,\phi(x_n)\right]$

Then the solution of (4) can be obtained through a linear system of equations [29]:

$$\begin{bmatrix} K+D_\alpha^{-1} & \mathbf{1} \\ \alpha^T K & \alpha^T \mathbf{1} \end{bmatrix} \begin{bmatrix} \beta^{(j)} \\ b^{(j)} \end{bmatrix} = \begin{bmatrix} y^{(j)} \\ \alpha^T y^{(j)} \end{bmatrix} \tag{6}$$

where $\alpha = \left[\alpha_1,...,\alpha_n\right]^T$, $y^{(j)} = \left[y_{1j},...,y_{nj}\right]^T$, $\mathbf{1} = (1,...,1)^T$, $D_\alpha = diag(\alpha_1,...,\alpha_n)$, $K = \boldsymbol{\varphi}\boldsymbol{\varphi}^T$ with $K_{ij} = k(x_i, x_j)$, where $k(x_i, x_j)$ is the kernel function. Here, we use radial basis function (RBF) to define the kernel function with $k(x_i, x_j) = \exp\left(-\|x_i - x\|_2^2 / 2\sigma^2\right)$. With the solution β^j of (6), we can get the regression model f , which can be achieved by an iterative procedure.

The procedure for solving (4) is summarized as follows:

Algorithm 1: The procedure for solving MSVR

Input: training data $(\mathrm{x}_i, \mathrm{y}_i)$ $\forall i = 1,...,n$, σ , ε , C .

Output: the regression model f .

(1) Initialization: β^j , b^j , $j = 1,...,k$.

(2) Compute corresponding α value by solving (5) according to KKT condition.

(3) Compute $\left[\beta^j, b^j\right]^T$ using (6) and $e_i = \left\|\mathrm{y}_i - \mathbf{W} \cdot \phi(\mathbf{x}_i) - \mathbf{b}\right\|$.

(4) Stop computing if all $e_i \le \varepsilon, i = 1, 2, ..., n$, otherwise go back to step 2.

C. Adaptive Sampling Policy

In this part, we present an adaptive packet sampling method specifically designed for multimedia traffic classification. This idea is based on the intuition that when some packets have the same PS and IAT values in the original set, the benefit of adding these packets to the sampled set for classification decreases and it's not necessary to select all these packets in the sampling process.

In order to predict the values of PS and IAT, we divide the time axis into a series of time slots with the same interval t_0 . We define $U(k) = \left\{P_1,...,P_{k_n}\right\}$ as the selected packets during the time interval $\left[k \cdot t_0, (k+1)t_0\right]$, where P_i denotes the i th packet, k_n is the number of the selected packets within this time interval. Let $\overline{\mathrm{PS}}, \overline{\mathrm{IAT}}$ respectively be the average value of the PS and IAT of the previous i packets during the time interval $\left[k \cdot t_0, (k+1)t_0\right]$, and $\mathrm{PS}_{pre}, \mathrm{IAT}_{pre}$ respectively be the predicted value of PS and IAT of the next packet P_{i+1}. The specific sampling idea is described as follows:

(1) If there exists a packet $P_j \in U(k), j \in [1,...,k_n]$ and the pair value $\left\langle \mathrm{PS}_j, \mathrm{IAT}_j \right\rangle$ meets such condition: $\mathrm{PS}_{pre} \in \left[\mathrm{PS}_j - \theta_{\mathrm{PS}}, \mathrm{PS}_j + \theta_{\mathrm{PS}}\right]$ and $\mathrm{IAT}_{pre} \in \left[\mathrm{IAT}_j - \theta_{\mathrm{IAT}}, \mathrm{IAT}_j + \theta_{\mathrm{IAT}}\right]$, which means that a packet with the same application type as the next predicted packet P_{i+1} is already selected with high probability, the sampling probability selecting the $(i+1)$ th packet should be reduced. We denote this condition by condition 1.

(2) If all the selected packets in $U(k)$ meet such condition: $\mathrm{PS}_{pre} \notin \left[\mathrm{PS}_j - \theta_{\mathrm{PS}}, \mathrm{PS}_j + \theta_{\mathrm{PS}}\right]$ and $\mathrm{IAT}_{pre} \notin \left[\mathrm{IAT}_j - \theta_{\mathrm{IAT}}, \mathrm{IAT}_j + \theta_{\mathrm{IAT}}\right]$, $j = 1,...,k_n$, which means that the next predicted packet P_{i+1} is of different application type with any packets in $U(k)$ with high probability, the sampling probability selecting the $(i+1)$ th packet should be enlarged. And we denote this condition by condition 2.

According to above sampling idea, we define the probability that the adaptive sampling will select the $(i+1)$ th packet as follows:

$$p(P_{i+1} \mid \mathrm{PS}, \mathrm{IAT}) = \begin{cases} \tau\upsilon & \text{condition 1 is met} \\ \tau\varsigma & \text{condition 2 is met} \end{cases} \tag{7}$$

where $\upsilon = \min\left(\frac{\log(\mathrm{PS}_{pre})}{\log(\overline{\mathrm{PS}})}, \frac{\log(\overline{\mathrm{PS}})}{\log(\mathrm{PS}_{pre})}, \frac{\log(\mathrm{IAT}_{pre})}{\log(\overline{\mathrm{IAT}})}, \frac{\log(\overline{\mathrm{IAT}})}{\log(\mathrm{IAT}_{pre})}\right)$

and $\varsigma = \max\left(\frac{\log(\mathrm{PS}_{pre})}{\log(\overline{\mathrm{PS}})}, \frac{\log(\overline{\mathrm{PS}})}{\log(\mathrm{PS}_{pre})}, \frac{\log(\mathrm{IAT}_{pre})}{\log(\overline{\mathrm{IAT}})}, \frac{\log(\overline{\mathrm{IAT}})}{\log(\mathrm{IAT}_{pre})}\right)$,

and $\tau \in [0,1]$ denotes the sampling rate. In the first equation of (7), condition 1 is satisfied, so the sampling probability decreases which is realized by using the term υ, the value of which is smaller than one. In the second equation of (7), condition 2 is satisfied, which means that there is a big deviation between the prediction value and historical value, so the sampling probability increases and it is realized by using the term ς, the value of which is bigger than one. It is noted that $p(P_{i+1} \mid \mathrm{PS}, \mathrm{IAT})$ is compelled to one if the value of $\tau\varsigma$ is bigger than one.

The four ratios of logs ($\frac{\log(\mathrm{PS}_{pre})}{\log(\overline{\mathrm{PS}})}, \frac{\log(\overline{\mathrm{PS}})}{\log(\mathrm{PS}_{pre})}, \frac{\log(\mathrm{IAT}_{pre})}{\log(\overline{\mathrm{IAT}})}, \frac{\log(\overline{\mathrm{IAT}})}{\log(\mathrm{IAT}_{pre})}$) in υ and ς are related to dynamic information of the traffic. And the choice of ratios of logs in υ and ς is only one of the feasible policies characterized by PS and IAT, and further study about how to obtain the best mechanism to decrease or increase the sampling probability will be done in future work.

The thresholds θ_{PS} and θ_{IAT} are the input parameters and they can be tuned dynamically to adapt the input data, which will also be discussed in our future work.

The proposed adaptive packet sampling is able to adjust the sampling probability for the ultimate goal of achieving representational dataset for classification. It selects packets according to PS and IAT in order both to suppress packets with the same application type, and to increase the sampled set with different values of

$\langle \mathrm{PS}, \mathrm{IAT} \rangle$ while keeping the numbers of samples within a tolerable range.

Since we have given the details the adaptive packet sampling method, the steps are as follows:

Algorithm 2: Adaptive packet sampling algorithm

Input: original finite set U' for training MSVR, and original finite set U to be sampled.

Output: the finite sampled set S from U.

(1) Constitution of the training data set. Let $X_i = (x_i, x_{i+1}, \cdots, x_{i+m-1})$ and $Y_i = X_{i+m}$, where $x_j \in U', j = i, ..., i+m$. Then we get the training data set $\{(X_i, Y_i)\}_{i=1}^{l}$ from U', where $l = |U'| - m$ with $|U'|$ denoting the number of packets in U' and m the embedding dimension.

(2) Using MSVR based on U', we get the final forecasting model f.

(3) For the predicted sample $x_j \in U$, the predicted value is $Y_j = f(x_{j-m}, x_{j-m+1}, \cdots, x_{j-1})$. In this way, we can get the predicted values of PS and IAT ($\mathrm{PS}_{pre}, \mathrm{IAT}_{pre}$) respectively.

(4) According to (7), the sampling probability of selecting the next packet is achieved. Then we can obtain the final sampling set S from U

D. Classification Algorithm

Since sampled traffic has rarely been used for multimedia traffic classification, it is still unclear which algorithm might be the most apt to deal with such data. Motivated by work such as Fan *et al.* [3], which proposes a VOVClassifier extracting features from frequency domain through two features: PS and IAT, our choice falls on two classification algorithms: VOVClassifier and SVM.

As machine learning is beyond the scope of this work, we refer the interested readers to Williams *et al.* [31] for a detailed description of different learning techniques, their merits and performance. At the same time, we need briefly to introduce VOVClassifier and SVM, as it will be instrumental to our experimental analysis later on. The VOVClassifier relies on two main characteristics of packets in voice and video flows: PS and IAT. The approach first models each flow into a two-dimensional stochastic process, and then uses the *Power Spectral Density* analysis to dig the hidden regularities constituting the fingerprint of the flow. The authors show that these fingerprints are unique for voice and video flows as well as each multimedia application that generates these flows, which can be easily clustered to create a voice and video subspace. These subspaces can be separated by a linear classifier.

As for SVM classifier, the initial form of SVM is a binary classifier where the output of the classifier is either positive or negative. A multi-class classification can be implemented by combining multiple binary classifiers using pairwise coupling method [32]. Binary SVM is a classifier which discriminates data points into two categories. Each data point is represented by a multi-dimensional vector. And each of these data points belongs to only one of the two classes. The overall aim is to achieve maximum separation between the two classes, which is obtained by introducing a separating hyperplane. This hyperplane must maximize the margin between the two classes, which is known as the optimum separating hyperplane. Such a hyperplane generalizes better very likely, meaning that the hyperplane can classify correctly the testing data points.

IV. Sampling Performance

To dig deeper into the analysis of the impact of sampling on traffic measurement, in this section we will evaluate the benefits of the proposed adaptive packet sampling technique. The verification is based on comparison of adaptive sampling with other sampling techniques on real network traffic data set. In this paper, we focus on three main categories of multimedia applications: VoIP, Instant Messaging (IM), and IPTV. We obtained the labeled real traffic traces for the three multimedia applications: VoIP (Skype), IM (MSN, Yahoo, GTalk), and IPTV, from the site [33]. The detailed information of the traffic traces is summarized in Table I. Here, we performed two types of experiments. First, we measured performance of each sampling method to show differences in computational complexity. Then we inspected the influence of the sampling methods on traffic feature distributions for the sampling loss and sampling loss coefficient. In the experimental process, we define t_0 =1min, m =10, θ_{PS} =10bytes, θ_{IAT} =1ms, $\sigma = 0.5$, $\varepsilon = 0.01$ and $C = 10$.

TABLE I: Application Breakdown of the Traffic Traces

Application	Traces	Duration	# packets	Date size(MB)
Skype	Skype1	95 hour 26 min	2357997	338.5
	Skype2	95 hour 45 min	39627543	8396.8
	Skype3	79 hour 3 min	3049284	231.3
IM	MSN	95 hour 45 min	15434573	2234.3
	YMSG	95 hour 45 min	841221	79.1
	XMPP	95 hour 45 min	214636	34.8
IPTV	IPTV	5 min 32 sec	13513514	18633.8

A. Computational Complexity

The performance of sampling techniques can be described in terms of CPU requirements. First of all, it is important to understand that the presented sampling technique involves three rather different procedures as seen in algorithm 1: MSVR model training, prediction operation for PS and IAT, and selecting packets adaptively. The first relates to the solution of an optimization problem solved by an iterative procedure. The second only involves a limited number of simple operations as can be gathered by using the regression function. And the last involves a comparison between the forecast values and the foregone information. These procedures are performed at intrinsically different timescales, and the first procedure will consume the longest time due to its iterative procedure.

The experiment was performed with two sizes of multimedia traffic on three sampling methods, namely random sampling, adaptive sampling predicting PS and IAT simultaneously (Adaptive S sampling), and adaptive sampling predicting PS and IAT respectively (Adaptive R sampling). The experimental result is shown in Table II, reporting the typical computational performance of our experimental campaign, which was run on a PC featuring a 2.7 GHz Pentium (R) Dual-Core processor equipped with 2 GB of RAM. The Adaptive S sampling, solving the programming problem (5) needing an iterative procedure, requires more computational time than the other methods as you can see in Table II. It is noted that the absolute values are not as important as the relative differences between the individual methods and sizes of input samples. Random sampling method (sampling probability is 0.1) requires minimal computational time, but it has the worst sampling performance as we will show further in next Subsection.

TABLE II: CPU TIME NEEDED TO SAMPLE INPUT PACKETS BY USING THREE TYPES OF SAMPLING METHODS (MS)

Raw Packets	Random sampling	Adaptive R	Adaptive S
300 000	771	1026	1978
8 400 000	13321	20801	28986

B. Sampling Distortion

To study the sampling effect on the number of packets with different values of $\langle \mathrm{PS}, \mathrm{IAT} \rangle$, we will compare the differences of these three packet sampling techniques (random sampling, Adaptive S sampling and Adaptive R sampling). In this process, we use 1-hour multimedia traffic for training data, and 5-minute multimedia traffic for prediction.

TABLE III: SAMPLING PERFORMANCE OF THREE TYPES OF SAMPLING METHODS, THE NUMBER OF RAW PACKETS FOR PREDICTION IS 520000 AND N_U =15672

Metrics	Random sampling	Adaptive R	Adaptive S
$d(U,S)$	6961	2070	351
$\Phi(U,S)$	55.6%	86.8%	97.8%

The impacts of above mentioned packet sampling methods on the sampling loss $d(U,S)$ and sampling loss coefficient $\Phi(U,S)$ are described in Table III. As you can see in Table III, Adaptive S method clearly outperforms other two sampling methods, which means that it preserves better the number of different packets. In this evaluation, we also demonstrated that the number of the sampled packets using the proposed approach is much more reversible than random sampling, which implies that it is a promising approach to provide a representative training dataset for the traffic classification methods. It also shows in Table III that the information content of these features (PS and IAT) is altered to a certain extent; however their information would still be extremely valuable for classification purposes which will be demonstrated in the next section.

V. CLASSIFICATION PERFORMANCE UNDER ADAPTIVE SAMPLING

After the analysis of computational complexity and the distortion due to sampling, in this section we provide a detailed evaluation of the impact of sampling strategies on the classification performance by using two machine learning methods, namely the methods called VOVClassifier and SVM. It is noteworthy that the training and validation data are gathered at the same sampling policy. To test the method's effectiveness, we adopt the harmonic mean F-score of *Recall* and *Precision* as the evaluating metric, which are calculated as below:

$$Recall = n/N \tag{8}$$

$$Precision == n/\bar{n} \tag{9}$$

$$\text{F-score} = \frac{2 * Recall * Precision}{(Recall + Precision)} \tag{10}$$

where N is the total number of samples in the test data set to be classified for each multimedia application(Skype, IM or IPTV); n is the total number of samples which are correctly classified by VOVClassifier or SVM; $\bar{n}$ is the total number of samples which are classified as some application type and $\bar{n}$ may include the wrongly classified samples. To estimate classification performance, we rely on 5-fold cross-validation: we partition the data set into five complementary subsets of equal size. Four subsets are used as training data; the remaining subset serves as test data. We repeat this process five times such that each of the five subsets is used exactly once as test data.

The whole evaluation results are shown in Fig. 2, Fig. 3 and Fig. 4. Fig. 2 describes the comparison of the classification performance between VOVClassifier and SVM without any sampling. And Fig. 3 shows that Adaptive R has comparable results with no sampling method. We can see that adaptive S significantly increases the classification performance when compared to the above two cases in Fig. 4, which means we can use Adaptive S without any significant damage to traffic classification effectiveness.

These results show that, even though the number of packets with different values of $\langle \mathrm{PS}, \mathrm{IAT} \rangle$ is affected, the information they convey on the application type is still important for the classification purpose. On the contrary, the unsampled set is unsuitable for identifying the real traffic although it could capture the properties of traffic well. We can also find that SVM has poorer performance than VOVClassifier, due to the fact that the SVM classifier does not make full use of the strong regularities between PS and IAT.

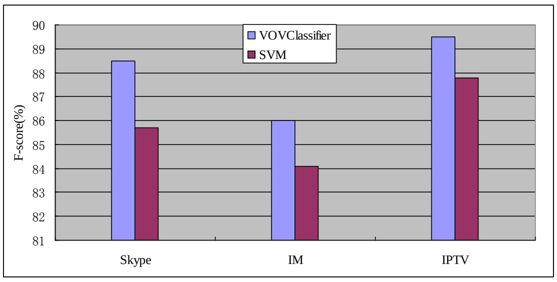

Fig. 2. F-score of VOVClassifier and SVM without sampling

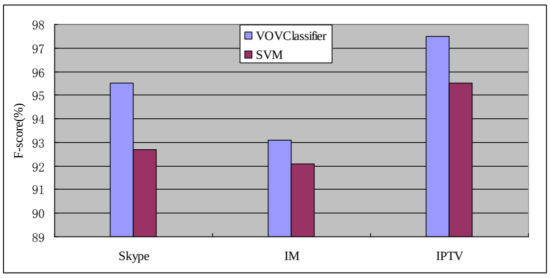

Fig. 3. F-score of VOVClassifier and SVM with Adaptive R sampling

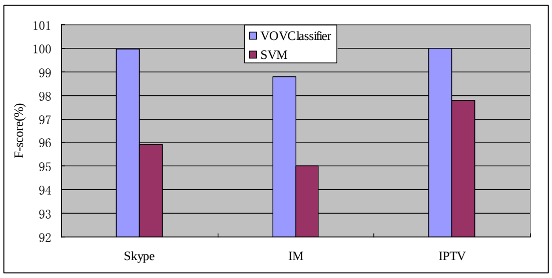

Fig. 4. F-score of VOVClassifier and SVM with Adaptive S sampling

VI. CONCLUSIONS

In this paper, we proposed an adaptive packet sampling method and then studied its impact on multimedia traffic classification performance. Packet sampling is already a very common practice in operational networks and the increasing trend of network traffic is likely to spread its adoption even more among network operators. For this reason, this paper presents an adaptive packet-based sampling method and assessed its sampling performance as well as the impact on the multimedia traffic classification.

This proposed packet sampling technique reduces the loss of information caused by sampling procedure, which suppresses large redundant packets with similar values of PS and IAT, while focusing on selecting the representative samples from the classification perspective. We evaluate different properties between adaptive and random sampling algorithms using two evaluating indexes and reveal their impact using two machine learning classification methods with real traffic. According to our experimental results, we showed that the presented adaptive sampling method, selecting packets according to PS and IAT predicted by using MSVR, is a promising approach that has good sampling performance and also is able to enhance the performance of the traffic classification methods.

In this adaptive packet sampling technique, we adjust the sampling probability according to only two features extracted from flows only designed for multimedia traffic. In the future, we will dig additional features used for adaptive sampling for classifying general traffic. Besides, we will deeper how to reduce the computational complexity of the adaptive sampling algorithm by means of optimizing iterative algorithms to make this sampling method more suitable for online traffic classification.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (NO. 60872043) and National 863 plans project (NO. 2009AA01Z207). We are grateful to the anonymous reviewers for their helpful comments. Also we want to thank Linlin Li for critically reviewing the manuscript.

REFERENCES

[1] S. Valenti, D. Rossi, M. Meo, M. Mellia, and P. Bermolen, “Accurate, fine-grained classification of P2P-TV applications by simply counting packets,” in *Proc. International in Traffic Measurement and Analysis (TMA) Workshop at IFIP Networking'09*, 2009, pp. 84–92.

[2] M. Canini, W. Li, M. Zadnik, and A. W. Moore, “Experience with highspeed automated application-identification for network-management,” in *Proc. 5th ACM/IEEE Symposium on Architectures for Networking and Communications Systems*, 2009, pp. 209-218.

[3] J. Fan, D. Wu, A. Nucci, R. Keralapura, and L. Gao, “Protocol oblivious classification of multimedia traffic,” *Security and Communication Networks*, vol. 4, pp. 357–371, 2009.

[4] Y. Chabchoub, C. Fricker, F. Guillemin, and P. Robert, “On the statistical characterization of flows in Internet traffic with application to sampling,” *Computer Communications*, vol. 33, pp. 103-112, 2010.

[5] J. Drobisz and K. J. Christensen, “Adaptive sampling methods to determine network traffic statistics including the Hurst parameter,” in *Proc. 8th IEEE LCN*, 1998, pp. 238-247.

[6] Y. Chabchoub, C. Fricker, F. Guillemin, and P. Robert, “Deterministic versus probabilistic packet sampling in the internet,” in *Managing Traffic Performance in Converged Networks*, vol. 4516, pp. 678-689, 2007.

[7] N. Duffield, C. Lund, and M. Thorup, “Properties and prediction of flow statistics from sampled packet streams,” in *Proc. 2nd ACM SIGCOMM Workshop on Internet Measurement*, 2002, pp. 159-171.

[8] E. A. Hernandez, M. C. Chidester, and A. D. George, “Adaptive sampling for network management,” *Journal of Network and Systems Management*, vol. 9, pp. 409–434, 2001.

[9] T. Zseby, “Deployment of sampling methods for SLA validation with non-intrusive measurements,” in *Proc. PAM*, 2002, pp. 25-26.

[10] K. Bartos and M. Rehak, “Towards efficient flow sampling technique for anomaly detection,” in *Proc. 4th International Conference on Traffic Monitoring and Analysis*, 2012, pp. 93-106.

[11] D. Brauckhoff, B. Tellenbach, A. Wagner, M. May, and A. Lakhina, “Impact of packet sampling on anomaly detection metrics,” in *Proc. 6th ACM SIGCOMM Conference on Internet Measurement*, 2006, pp. 159-164.

[12] I. Paredes-Oliva, P. Barlet-Ros, and J. Solé-Pareta, “Portscan detection with sampled netflow,” *Traffic Measurement and Analysis*, vol. 5537, pp. 26–33, 2009.

[13] H. Jiang, A. W. Moore, Z. Ge, S. Jin, and J. Wang, "Lightweight application classification for network management," in *Proc. SIGCOMM Workshop on Internet Network Management*, 2007, pp. 299-304.
[14] V. Carela-Español, P. Barlet-Ros, A. Cabellos-Aparicio, and J. Sol-Pareta, "Analysis of the impact of sampling on netflow traffic classification," *Computer Networks*, vol. 55, no. 5, pp. 1083–1099, 2011.
[15] D. Tammaro, S. Valenti, D. Rossi, and A. Pescape, "Exploiting packet-sampling measurements for traffic characterization and classificayion," *International Journal of Network Management*, vol. 22, pp. 451–476, 2012.
[16] M. Zhang, W. John, K. Claffy, and N. Brownlee, "State of the art in traffic classification: A research review," in *Proc. 10th International Conference on Passive and Active Measurement, Student Workshop*, 2009, pp. 1–2.
[17] H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee, "Internet traffic classification demystified: myths, caveats, and the best practices," in *Proc. ACM CoNEXT Conference*, 2008, pp. 1–12.
[18] T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE Communications Surveys& Tutorials*, vol. 10, pp. 56–76, 2008.
[19] P. Amer and L. Cassel, "Management of sampled real-time network measurements," in *Proc. 14th Conference on Local Computer Networks*, 1989, pp. 62–68.
[20] K. C. Claffy, G. C. Polyzos, and H. W. Braun, "Application of sampling methodologies to network traffic characterization," *SIGCOMM Computer Communication Review*, vol. 23, pp. 194–203, 1993.
[21] N. Duffield, "Sampling for passive internet measurement: A review," *Statistical Science*, vol. 19, pp. 472–498, 2004.
[22] IPFIX. Internet engineering task force, IP flow information export. [Online]. Available: http://www.ietf.org/html.charters/ipfix-charter.html
[23] PSAMP. Internet engineering task force packet sampling working group. [Online]. Available: https://ops.ietf.org/lists/psamp
[24] N. Hohn and D. Veitch, "Inverting sampled traffic," in *Proc. 3rd ACM SIGCOMM Conference on Internet Measurement*, 2006, pp. 222–233.
[25] C. Estan and G. Varghese, "New directions in traffic measurement and accounting: Focusing on the elephants, ignoring the mice," *ACM Transactions on Computer Systems*. vol. 21, pp. 270–313, 2003.
[26] B. Choi, J. Park, and Z. Zhang, "Adaptive random sampling for load change detection," in *Proc. ACM SIGMETRICS '02*, 2002, pp. 272-273.
[27] J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson, "Offline/realtime traffic classification using semi-supervised learning," *Performance Evaluation*, vol. 64, pp. 1194–1213, 2007.
[28] J. Park, H. R. Tyan, and C. C. Kuo, "Internet traffic classification for scalable QoS provision," in *Proc. IEEE International Conference on Multimedia and Expo*, 2006, pp. 1221–1224.
[29] F. Perez-Cruz, G. Camps-Vails, and E. Soria-Olivas, "Multi-dimensional function approximation and regression estimation," in *Proc. International Conference on Artificial Neural Networks*, 2002, pp. 757-762.
[30] B. Schoelkopf and A. Smola, *Learning with Kernels*, M.I.T. Press, 2001.
[31] N. Williams, S. Zander, and G. Armitrage, "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification," *ACM SIGCOMM Computer Communication Review*, vol. 36, pp. 5–16, 2006.
[32] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *Advances in Neural Information Processing Systems*, vol. 26, pp. 451-800, 1998.
[33] Atstat Traces. (2012). [Online]. Available: http://tstat.tlc.polito.it/traces.shtml

Rongqiang Lin was born in Fujian Province, China, in 1990, master candidate. His main research interests include computer network, information theory and communication technology.

Ou Li was born in Henan Province, China, in 1961, professor, Ph. D. supervisor. His main research interests include computer communication network, Cognitive Radio Network, Wireless sensor network and ad-hoc network

Qing Li was born in Hebei Province, China, in 1972, professor, Ph. D. supervisor. Her main research interests include computer network and information security, and information operations

Kun Dai was born in Henan Province, China, in 1982, Ph. D. candidate. His main research interests include computer network, network security and communication technology