Journal of Computer and System Sciences 80 (2014) 1021–1036

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Internet traffic clustering with side information

Yu Wang [a], Yang Xiang [a,*], Jun Zhang [a], Wanlei Zhou [a], Bailin Xie [b]

[a] School of Information Technology, Deakin University, Melbourne, Australia
[b] Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou, China

ARTICLE INFO

Article history:
Received 25 September 2012
Received in revised form 15 March 2013
Accepted 27 August 2013
Available online 12 February 2014

Keywords:
Traffic classification
Semi-supervised machine learning
Constrained clustering

ABSTRACT

Internet traffic classification is a critical and essential functionality for network management and security systems. Due to the limitations of traditional port-based and payload-based classification approaches, the past several years have seen extensive research on utilizing machine learning techniques to classify Internet traffic based on packet and flow level characteristics. For the purpose of learning from unlabeled traffic data, some classic clustering methods have been applied in previous studies but the reported accuracy results are unsatisfactory. In this paper, we propose a semi-supervised approach for accurate Internet traffic clustering, which is motivated by the observation of widely existing partial equivalence relationships among Internet traffic flows. In particular, we formulate the problem using a Gaussian Mixture Model (GMM) with set-based equivalence constraint and propose a constrained Expectation Maximization (EM) algorithm for clustering. Experiments with real-world packet traces show that the proposed approach can significantly improve the quality of resultant traffic clusters.

1. Introduction

Internet traffic classification is the process of identifying network applications and classifying the traffic accordingly. It plays an important role in modern network management and security systems and in cloud computing environments. By obtaining full visibility into the types of traffic traversing through the network, accurate traffic classification enables fine-grained management of application performance and balanced utilization of the network resources.

Traditional traffic classifiers are port-based, which simply inspect the transport layer port number fields in packet headers and predict the upper layer application according to the list of well-known and registered port numbers maintained by Internet Assigned Numbers Authority (IANA). This simple approach is efficient since it only requires access to packet headers. However, it has become increasingly inaccurate in recent years due to the violation of port number assignments by more and more newly emerging applications. The alternative approach widely deployed in industry is payload-based, which performs deep packet inspection to either reconstruct and validate application protocol sessions or match protocol signatures against payload contents. These methods are accurate, but they not only impose significantly higher computational complexity, but also require specific knowledge of the target application protocols in advance. The a priori information, such as message

* Corresponding author.
E-mail addresses: y.wang@deakin.edu.au (Y. Wang), yang@deakin.edu.au (Y. Xiang), jun.zhang@deakin.edu.au (J. Zhang), wanlei@deakin.edu.au (W. Zhou), xiebailin96@126.com (B. Xie).

http://dx.doi.org/10.1016/j.jcss.2014.02.008
0022-0000/

formats and protocol signatures, is typically derived from protocol specifications or packet traces manually by network experts and the knowledge database has to be updated from time to time. This means that the approach involves heavy human intervention and cannot handle unknown applications automatically. Moreover, the classifiers fail when the access to payload content is unavailable, such as the case of classifying encrypted traffic.

Due to the limitations of the classic port-based and payload-based approaches, recent research efforts have been dedicated to developing alternative classification schemes. One of the major directions is to exploit packet and flow level characteristics, which capture the distinctive usage and interaction patterns of network applications, typically by applying machine learning (ML) techniques [1]. In this approach, a vector of features is extracted to describe each traffic flow, where the elements are some specific flow statistics such as the mean and variance of packet sizes or inter-packet times. The feature vectors are supplied as input to the machine-learning engine, in which two separate stages are involved. The first stage is offline learning where a classifier (e.g. a probabilistic model or a set of classification rules) is learned from a set of training data, and the second stage is online classifying in which the classifier is used to predict the application class of real-time traffic.

Depending on whether or not the training data set is fully labeled, two general types of learning methods are applicable: supervised learning (or classification) and unsupervised learning (or clustering). In supervised learning the training data is fully labeled and the goal is to find a mapping from input features to output classes. In contrast, unsupervised learning focuses on discovering patterns in unlabeled training data such that the flows with similar characteristics are grouped into clusters without any prior guidance from class labels. The resultant traffic clusters need to be labeled and transformed into a classifier for the online classifying stage. Clustering techniques are important since in practice it is very difficult and labor intensive to obtain a fully labeled data set. Besides, clustering has the potential to discover novel patterns that represent previously unknown applications. However, previous work [2–9] has shown that the application of some classic clustering algorithms, such as K-Means and EM, yielded relatively low accuracy.

In this work, we present a novel semi-supervised traffic clustering approach that generates highly accurate traffic clusters. The motivation is that there exists abundant side information describing partial equivalence relationships across flows in addition to the flow feature observations. For example, all concurrent flows that are connecting to the same destination IP address at the same port (i.e. the same endpoint) are typically using the same network application. This kind of side information can be efficiently derived by observing the 3-tuples of {Destination IP, Destination Port, Protocol} without knowing the actual class labels of the flows, and it can provide valuable guidance in the task of traffic clustering. In order to take advantage of the side information, we introduce the notion of set-based instance level constraint, which specifies that a particular set of equivalence flows have to be put in the same cluster during the clustering procedure. We then formulate the constrained clustering problem with the classic Gaussian Mixture Model (GMM) and apply a variant of the Expectation Maximization (EM) algorithm to fit the model with both the observed data and the equivalence constraints. Moreover, we also investigate the impact of unsupervised feature discretization for traffic clustering. The rationale is that some of the flow features are discrete in essence but they were measured and treated as continuous values in most of the related work. In particular, we use equal-frequency binning to quantize the numeric feature values.

To evaluate the proposed method, we use four real-world Internet traffic data sets taken from three locations around the world, including a packet trace captured in 2006 at the Internet edge of a university campus network, two collected in 2008 and 2009 at a trans-Pacific backbone link and the final trace recorded in a commercial ISP network in Australia in 2010. All of the data sets have either partial or full payload content that allows us to build the ground truth with high confidence. Several findings are made from the experimental results. First, the side information in terms of flow-level equivalence constraints is widely available in network traffic. Over 96.8% of flows across the data sets involve in the constraints. Second, an improvement of up to 8.5% in terms of overall accuracy can be achieved by incorporating the constraints into the clustering procedure. Third, by performing unsupervised discretization on features before clustering we can achieve a further improvement of up to 8.4% of flow accuracy. In short, the proposed approach significantly boosts the accuracy performance of Internet traffic clustering by incorporating the side information and performing feature discretization.

The remaining of the paper is organized as follows. A brief review of the related work on statistics-based traffic classification is presented in the next section. In Section 3, we discuss the proposed semi-supervised approach for Internet traffic clustering. The data sets used in the evaluations are described in details in Section 4, where an analysis of the side information in real-world traffic is also presented. Next is the experimental results given in Section 5. In Section 6, we discuss some practical issues of applying the semi-supervised traffic clustering approach in operational networks. Section 7 concludes this paper.

2. Related work

Clustering techniques have been applied in the context of Internet traffic analysis for a long time. In [2] Hernández-Campos et al. proposed using the triplets of (request data size, response data size, quiet time) for application-level communication modeling, such that taxonomy of important patterns in the traffic mix could be developed using hierarchical clustering methods. Similarly, McGregor et al. applied in [3] the Expectation Maximization (EM) algorithm for clustering packet traces based on a range of flow-level statistical attributes such as packet length and inter-arrival statistics, byte count and connection duration. The resulted clusters were then visualized and interpreted using radar charts. Although