Neurocomputing 168 (2015) 365–381

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A class-oriented feature selection approach for multi-class imbalanced network traffic datasets based on local and global metrics fusion

Zhen Liu [a], Ruoyu Wang [b,c,*], Ming Tao [c,d], Xianfa Cai [a]

[a] School of Medical Information Engineering, Guangdong Pharmaceutical University, Guangzhou 510006, China
[b] Information and Network Engineering and Research Center, South China University of Technology, Guangzhou 510006, China
[c] School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
[d] School of Computer, Dongguan University of Technology, Dongguan 523808, China

ARTICLE INFO

Article history:
Received 7 January 2015
Received in revised form
6 April 2015
Accepted 25 May 2015
Communicated by: Liang Lin
Available online 7 June 2015

Keywords:
Feature selection
Local metrics
Multi-class imbalance
Data drift
Network traffic

ABSTRACT

Feature selection is often used as a pre-processing step for machine learning based network traffic classification. Many feature selection techniques have been developed to find an optimal subset of relevant features and to improve overall classification accuracy. But such techniques ignore the class imbalance problem encountered in network traffic classification. The selected feature subset may bias towards the traffic class that occupies the majority of traffic flows on the Internet. To address this issue, this paper proposes a new approach, called class-oriented feature selection (COFS), to identify a relevant feature subset for every class. It combines the proposed local metric and the existing global metric to yield a potentially optimal feature subset for each class, and then removes the redundant features in each feature subset based on the *weighted symmetric uncertainty*. Additionally, to enhance the generalization on network traffic data, an ensemble learning based scheme is presented with COFS to overcome the negative impacts of the data drift on a traffic classifier. Experiments on real-world network traffic data show that COFS outperforms existing feature selection techniques in most cases. Moreover, our approach achieves $>96\%$ flow accuracy and $>93\%$ byte accuracy on average.

1. Introduction

A lot of network applications run on the Internet and generate a large amount of traffic every day [1]. It is a daunting task to manage the huge and mixture traffic. Network traffic classification is more and more important along with the development of Internet [2,3]. It aims to map the IP packets into a network application, so as to induct the network management activities, including supervisory control and data access [4], accounting and bandwidth consumption management etc. [5]. A variety of network traffic classification techniques have emerged [6]. Previous techniques are based on port numbers and payload information [7]. However, these did not work well on the traffic generated by the applications with dynamic port numbers and encrypted payloads. Payload based traffic classification techniques [8] rely on deep inspection of packet content, which obviously result in significant number of processing and memory requirements.

Recently, machine learning techniques are applied in network traffic classification [3,9,10]. These techniques can deal with above limitations by avoiding deep packet inspection and creating new features from transport layer statistics (e.g., packet size and interarrival time) [11–14]. A variety of features have been extracted from traffic flows in recent years. However, the presence of irrelevant or redundant features degrades the classification accuracy, maximizes the training and testing processing time of the classifier, and finally increases its storage requirements [15]. With the object of searching a subset of good features, feature selection techniques can play an effective role in removing irrelevant and redundant features.

1.1. Problem statement

Some classes (also called majority classes, e.g., WWW) often own a large number of flows but others (also called minority classes, e.g., ATTACK) only a small number of flows on a multi-class traffic dataset [16]. One of key challenges (for traffic classification) is the multi-class imbalance problem. Many feature selection techniques [17,18] have been developed to find an optimal feature subset and to improve overall classification accuracy. However, the selected features often perform well on majority classes and degrade the predictive accuracy of the classifier for minority classes, since the accuracies of majority classes contribute more to the overall classification accuracy.

* Corresponding author at: Information and Network Engineering and Research Center, South China University of Technology, Guangzhou 510006, China.
Tel.: +86 02087110596.
E-mail address: rywang@scut.edu.cn (R. Wang).

http://dx.doi.org/10.1016/j.neucom.2015.05.089
0925-2312/© 2015 Elsevier B.V. All rights reserved.

Most feature selection techniques [17–19] search only one optimal feature subset for the whole training set. But a feature subset may be not optimal for every class. For example, a feature subset is good for classifying WWW flows but may be bad for classifying ATTACK flows especially when ignoring the class imbalance problem. And its scalability is limited. When a new network application emerges, the whole process of feature selection has to be implemented again, requiring more time and effort. In addition, a network application class has a unique communication behavior and requires a specific feature subset to discriminate its traffic from others. Thereby, another key challenge (in feature selection) is the lack of algorithm to search a relevant feature subset for each application class on network traffic data.

Internet traffic data are very huge. The collected traffic datasets are just a small part of all traffic in the world. The traffic data are also dynamic. The features are subject to change depending on the monitored site. The other key challenge (for classification) is the data drift problem.

1.2. Contributions

This paper deals with the problems described above. It proposes a novel feature selection algorithm as well as a learning scheme to enhance the capabilities of network traffic classification task. Main contributions are as follows.

1) A new algorithm called class-oriented feature selection (COFS) is proposed to search an optimal feature subset for every class. With this object, a local metric is devised to assess the local correlation between a feature and a class. It is based on *maximum entropy*. COFS proceeds in two phases. First phase evaluates the local score and global score of each feature. The local score is the value of local metric while the global score is the value of global metric. And then it searches a relevant and discriminative feature subset for every class by comparing the local and global scores with cut-off thresholds. Instead of a fixed threshold, COFS individually solves an adaptive threshold for each local score when searching the locally relevant features for a class, so as to automatically suit the feature distribution. In the second phase, all selected feature subsets are passed to the procedure of removing redundant features. This procedure outputs an optimal feature subset for every class that will be used for network traffic classification.
2) A learning scheme is proposed within the ensemble learning framework to meet the context of COFS and improve the generalization on network traffic data. It also proceeds in two phases at the base of the feature subsets gained by COFS. In the first phase, the origin training set is projected into each feature subset and each reduced training set is individually used to train a classifier. In the second phase, a weight vector is resolved and will be assigned to the predication distribution of each classifier, aiming at reinforcing the strength of a classifier. Finally, an ensemble classifier is built.
3) A series of experiments are carried out on real-world network traffic datasets, in order to evaluate the performance of our approach, including the classification accuracy, efficiency and robustness. We compare our approach against three techniques, BFS (balanced feature selection) [27], WSU_AUC (weighted symmetric uncertainty_area under ROC curve) [32] and GOA (global optimization approach) [15], presented recently for network traffic classification task. Classification results show that COFS indeed improves our previous work BFS. COFS consumes much less time than WSU_AUC and GOA, and the efficiency is improved by a factor of 129%. COFS is also compared with a classical over sampling method SMOTE (Synthetic Minority Over-sampling Technique) on handling multi-class imbalance problem in network traffic classification. In overall, network traffic classification results show that COFS obtains higher flow *g*-mean and byte *g*-mean in most cases, and acquires higher flow precisions and byte precisions for most classes.

1.3. Organization

The rest of this paper is organized as follows. Section 2 reviews the related work of flow feature extraction and feature selection methods in network traffic classification. Section 3 presents our feature selection algorithm and learning scheme. Section 4 introduces the experimental datasets and performance evaluation metrics. Section 5 carries out experiments to validate the effectiveness of our approach in the way of comparing it against existing works on real-world network traffic datasets. Section 6 concludes this paper.

2. Related work

2.1. Related work on feature extraction

The “feature extraction” here is different from that in machine learning field. In this paper, it means exploiting new features to characterize traffic flows. A part of typical flow features is summarized in Table 1, where bulk transfer mode is defined as the time when there are more than three successive packets in the

Table 1
Summary of traffic flow features.

References	Traffic flow features
McGregor et al. [20]	Packet size statistics (minimum, maximum, quartiles, minimum as fraction of max and the first five modes); interarrival time (ITA); total number of bytes; flow duration; the number of transitions between transaction mode and bulk transfer mode; idle time in bulk mode
Roughan et al. [21]	Packet level: average/variance/root mean square of packet sizes Flow level: average of flow duration/volume/packets, variance of packets Connection level: the symmetry of a connection, advertised window sizes and the throughput distribution Intraflow/Connection: interarrival time, loss rates, latencies
Zander et al. [22]	Mean/variance of packet sizes in forward/backward direction; mean/variance of ITA in forward/backward direction; total number of bytes; flow duration
Moore et al. [12]	Port number (service, client); ITA (minimum, media, mean, first quartile, etc.); packet size (minimum, media, mean, first quartile, etc.); the number of packets (ACK, PUSH, etc.) Effective Bandwidth based upon entropy; FFT of packet IAT; Flow duration
Jiang et al. [23]	Traffic volume; distribution of traffic volume in terms of the time, space and applications Ratios of upload and download traffic
Bermolen et al. [25]	The number of peers that sent $(2^{Bn-1}+1, 2^{Bn})$ packets to a specific peer