Computers and Electrical Engineering 000 (2018) 1–8

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Network traffic classification based on transfer learning☆

Guanglu Sun[a,b], Lili Liang[a], Teng Chen[a], Feng Xiao[a], Fei Lang[b,c,*]

[a] School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, 150080, China
[b] Research Center of Information Security & Intelligent Technology, Harbin University of Science and Technology, Harbin, 150080, China
[c] School of Foreign Languages, Harbin University of Science and Technology, Harbin, 150080, China

ARTICLE INFO

Article history:
Received 3 September 2017
Revised 4 March 2018
Accepted 5 March 2018
Available online xxx

Keyword:
Traffic classification
Machine learning
Transfer learning
Domain adaptation
Maximum entropy model

ABSTRACT

Machine learning models used in traffic classification make the assumption that the training data and test data have independent identical distributions. However, this assumption might be violated in practical traffic classification due to changes of traffic features. The models trained by existing data will be ineffective in classifying new traffic. A transfer learning model without making the above assumption is proposed in the present study. The maximum entropy model (Maxent) was adopted as the base classifier in the transfer learning model. To examine the efficacy of the proposed method, the traffic dataset collected at the University of Cambridge was used in the condition that the training and test dataset were not identical. Experimental results showed that good classification performance was obtained based on the transfer learning model.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

With the increasing number of threats to networks, it is critical for network managers to have deeper understanding about the applications running in their networks. Traffic classification is a vital aspect of network management which enables to make classification of all network traffic [1].

In the last few years, many new network protocols tried to escape from monitoring by means of the disguised methods, such as dynamic port, encapsulation and encryption. It rendered port-based and payload-based methods unreliable. The researchers were motivated to use the application types as categories, the traffic statistical properties generated by network communication as features. Then machine learning models were utilized. It is widely used for machine learning based methods in traffic classification, including supervised learning model, unsupervised learning model and semi-supervised learning model [2]. However, there are still two big challenges in traffic classification. The first is being in line with the rapid growth of new applications. The second is that different training models are required as network topology and time change. In unsupervised learning model, it is hard to build a practical traffic classifier with the clustering results and no regard for the instruction of the real traffic classes [3]. As for semi-supervised learning model, it utilizes a small set of labeled traffic flows and a large set of unlabeled traffic flows during the training process. And for supervised learning model, a bigger set of labeled traffic flows are applied. Both of them achieved satisfying results, when the training and test dataset are identical. However, the changes of time, locations and traffic types make traditional machine learning models less effective than expected, because the training dataset collected in the past and the newly generated test data are not identical [4]. Although the model trained by the outdated data, cannot work well on the new network traffic, the outdated data should not be given

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. S. Liu.

* Corresponding author.
E-mail address: langfei@hrbust.edu.cn (F. Lang).

https://doi.org/10.1016/j.compeleceng.2018.03.005
0045-7906/© 2018 Elsevier Ltd. All rights reserved.

Please cite this article as: G. Sun et al., Network traffic classification based on transfer learning, Computers and Electrical Engineering (2018), https://doi.org/10.1016/j.compeleceng.2018.03.005

up, because it is difficult to acquire the gold-standard labeled traffic, which has the ability to describe the whole network traffic for machine learning methods. Thus, the outdated data with valuable knowledge should be reused for new tasks of traffic classification. It is unnecessary for transfer learning to be with the independent identical distribution (iid). And even for different tasks, transfer learning still has advantages to overcome the above obstacles [5]. It is assumed that a learning task and corresponding data are in a source domain, and a learning task and corresponding data are in a target domain, the objective of transfer learning is to improve rate of target prediction functions by learning the knowledge from the source domain. In contrast to traditional machine learning models, transfer learning does not need to make the iid assumption on the data between two domains [6].

Considering the situation that the data distribution will become different with the change of time, locations and traffic types, TrAdaBoost model [6] are introduced for traffic classification task, which is a distinguished inductive transfer learning model based on instance. Maxent is meanwhile utilized as a base classifier. Aiming at transferring the valuable knowledge in the source domain to the target task, TrAdaBoost uses few labeled data from the target domain to evaluate the availability of the data in the source domain. After that, the valuable auxiliary data are extracted from the source data and combined with the above labeled data in the target domain to train the classifier. By means of transferring the useful knowledge from the source domain to the target domain, TrAdaBoost helps the learning task in the target domain. The key contributions made by the present study are shown as the following.

- In the new traffic classification task, TrAdaBoost is presented to utilize the labeled traffic data extracted from different network traffic sources.
- Maxent model is applied as a base classifier in TrAdaBoost. The proposed method implements the transfer of traffic knowledge from the source domain to the target domain.
- The experiments are conducted to evaluate the performance of TrAdaBoost. The proposed method is compared with traditional machine learning methods, in the condition that there is no enough labeled data to train a learning model effectively in the target domain.

The structure of the present study is as the following. Section 2 briefly introduces related work of traffic classification and transfer learning. Section 3 summarizes the notations and tasks definition. Section 4 proposes the algorithms in detail. Then the datasets and the experimental results are given in Section 5. Section 6 concludes the present study and discusses future work.

2. Related work

In the past decade, a vast outpouring of research is carried out on various methods to deal with internet traffic classification, such as port-based methods, payload-based methods, behavior-based methods and machine learning based methods. Most of the methods are summarized in several surveys chronologically [1,7]. Among these methods, machine learning methods have been paid increasing attention to for its good performance in traffic classification task. Nguyen and Armitage reviewed the related studies based on machine learning models till 2008 in detail [2]. Thus, we mainly review the previous work of machine learning based methods after 2008.

The machine learning methods used for traffic classification task are classified into supervised learning methods, semi-supervised learning methods and unsupervised learning methods. In supervised learning methods, the traffic flows are manually labeled according to the generating application categories, as the gold-standard benchmark dataset. The supervised machine learning model is built based on the labeled flows, and its features are statistical patterns extracted from the flows. The new traffic flows are classified by the model that adjusts its parameters during the training process. In this way, many models were implemented in traffic classification. Moore and Zeuv [8] put forward Bayesian method to identify application protocols. They further made the accuracy higher by refining the variants. Five supervised models were compared in their accuracy of classification and performance of computation, which includes Naive Bayes with discretization, Naive Bayes with kernel density estimation, C4.5 decision tree, Bayesian network and naive Bayes tree [9]. Este et al. [10] applied Support Vector Machines model (SVMs) on three types of well-known datasets and obtained an average accuracy over 95%, 2.3% over the best performance of Bayesian methods and other methods on the same datasets. Finamore et al. [11] further presented statistical characterization of payload as features and used SVM to conduct traffic classification. Nguyen et al. [12] trained the ML models with a set of sub-flows and investigated various strategies of sub-flow selection. The accuracy of their model would be maintained when the traffic mixed up bi-directional flows. A hybrid method with heuristic rules and REPTree model was proposed to classify P2P traffic with different levels of features [13]. Li et al. [14] utilized logistic regression model to classify the flows using non-convex multi-task feature selection. They tried a Capped $-\ell_1,\ell_1$ to learn the features of flows as the regularizer. Peng et al. [15] verified 5–7 packets are the best numbers of packets for early step traffic classification based on 11 well-known supervised learning models.

As for unsupervised methods, the main application is clustering-based methods. The clustering-based model automatically groups unlabeled traffic flows into a set of clusters. The clusters mapping to the different applications are utilized to train a new traffic model. The expectation maximization model was utilized to cluster traffic flows and label each cluster to an application manually [16]. The k-means, DBSCAN and AutoClass models were verified and summarized that the clustering methods got the good-quality clusters when the number of clusters reached a certain scale [17]. The method derived signatures from unidentified traffic flows automatically. However, the difference between the number of clusters and applications

Please cite this article as: G. Sun et al., Network traffic classification based on transfer learning, Computers and Electrical Engineering (2018), https://doi.org/10.1016/j.compeleceng.2018.03.005

Download English Version:

https://daneshyari.com/en/article/6883346

Download Persian Version:

https://daneshyari.com/article/6883346

Daneshyari.com