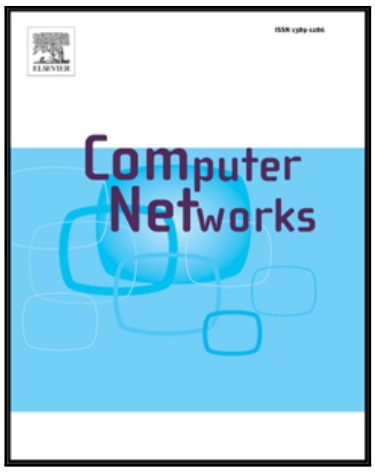

An Efficient Feature Generation Approach based on Deep Learning and Feature Selection Techniques for traffic classification

Hongtao Shi , Hongping Li , Dan Zhang , Chaqiu Cheng , Xuanxuan Cao

PII: S1389-1286(18)30008-2
DOI: 10.1016/j.comnet.2018.01.007
Reference: COMPNW 6355

To appear in: *Computer Networks*

Received date: 27 February 2017
Revised date: 1 November 2017
Accepted date: 9 January 2018

Please cite this article as: Hongtao Shi , Hongping Li , Dan Zhang , Chaqiu Cheng , Xuanxuan Cao , An Efficient Feature Generation Approach based on Deep Learning and Feature Selection Techniques for traffic classification, *Computer Networks* (2018), doi: 10.1016/j.comnet.2018.01.007

An Efficient Feature Generation Approach based on Deep Learning and Feature Selection Techniques for traffic classification

Hongtao Shi[a,b]*, Hongping Li[b], Dan Zhang[b], Chaqiu Cheng[b], Xuanxuan Cao[b]

[a] *Network Management Center, Qingdao Agricultural University, Qingdao 266109, PR China*
[b] *College of Information Science and Engineering, Ocean University of China, Qingdao 266100, PR China*

Keywords: Feature selection; Deep learning; Multi-class imbalance; Concept drift; Machine Learning; Traffic classification

Abstract: Substantial recent efforts have been made on the application of Machine Learning (ML) techniques to flow statistical features for traffic classification. However, the classification performance of ML techniques is severely degraded due to the high dimensionality and redundancy of flow statistical features, the imbalance in the number of traffic flows and concept drift of Internet traffic. With the aim of comprehensively solving these problems, this paper proposes a new feature optimization approach based on deep learning and Feature Selection (FS) techniques to provide the optimal and robust features for traffic classification. Firstly, symmetric uncertainty is exploited to remove the irrelevant features in network traffic data sets, then a feature generation model based on deep learning is applied to these relevant features for dimensionality reduction and feature generation, finally Weighted Symmetric Uncertainty (WSU) is exploited to select the optimal features by removing the redundant ones. Based on real traffic traces, experimental results show that the proposed approach can not only efficiently reduce the dimension of feature space, but also overcome the negative impacts of multi-class imbalance and concept drift problems on ML techniques. Furthermore, compared with the approaches used in the previous works, the proposed approach achieves the best classification performance and relatively higher runtime performance.

1. Introduction

Accurate classification of Internet traffic is the basis of many network management tasks [1, 2], including Quality of Service (QoS) control, intrusion detection and diagnostic monitoring. Traditional traffic classification approaches are based on examining the 16-bit port numbers in transport layer header or investigating the signature information in the packet payloads [3]. These approaches proved to be inefficient as they encounter many problems such as dynamic port numbers, data encryption and user privacy protection.

Due to the limitations of traditional traffic classification approaches, many research papers [4-11] have been dedicated to conduct traffic classification by applying ML techniques to flow statistical features. Although they made significant achievements, classifying Internet traffic by using ML techniques is still a daunting task, as the high redundancy of flow statistical features greatly degrades the accuracy and efficiency of the ML classifiers [12]. With the aim of solving this problem, FS techniques [13] can play an effective role in reducing the dimensionality (of flow statistical features) and removing irrelevant and redundant features. However, despite a vast number FS methods proposed in the literature [1,14,15,16], searching for the optimal features by FS methods remains a challenge because: (1) FS techniques conduct the search for an optimal subset using different evaluation criteria, which may make the optimal subset be local optima; (2) most FS techniques have been developed for improving classification accuracy by removing the redundant features, but neglect the stability of optimal subset for variations in the traffic data; (3) FS techniques cannot capture the complex dependency across all flow statistical features, which have a great impact on traffic classification. Thus, one of the key challenges is to provide the optimal and robust features for traffic classification.