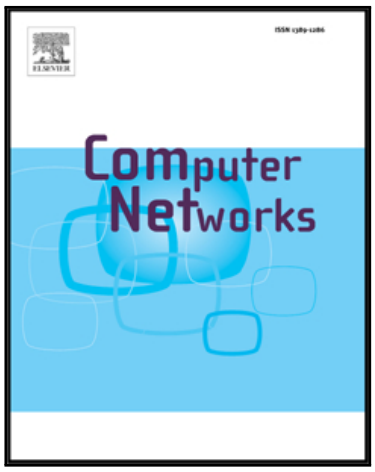

Efficient and robust feature extraction and selection for traffic classification

Hongtao Shi , Hongping Li , Dan Zhang , Chaqiu Cheng , Wei Wu

PII: S1389-1286(17)30078-6
DOI: 10.1016/j.comnet.2017.03.011
Reference: COMPNW 6126

To appear in: *Computer Networks*

Received date: 6 July 2016
Revised date: 10 February 2017
Accepted date: 10 March 2017

Please cite this article as: Hongtao Shi , Hongping Li , Dan Zhang , Chaqiu Cheng , Wei Wu , Efficient and robust feature extraction and selection for traffic classification, *Computer Networks* (2017), doi: 10.1016/j.comnet.2017.03.011

Efficient and robust feature extraction and selection for traffic classification

Hongtao Shi[a,b], Hongping Li[a]*, Dan Zhang[a], Chaqiu Cheng[a], Wei Wu[a]

[a] *College of Information Science and Engineering, Ocean University of China, Qingdao 266100, PR China*

[b] *Network Management Center, Qingdao Agricultural University, Qingdao 266109, PR China*

Keywords: Multifractal features; Wavelet leaders; Feature selection; Traffic classification; Optimal features; Machine learning

Abstract: Given the limitations of traditional classification methods based on port number and payload inspection, a large number of studies have focused on developing classification approaches that use Transport Layer Statistics (TLS) features and Machine Learning (ML) techniques. However, classifying Internet traffic data using these approaches is still a difficult task because (1) TLS features are not very robust for traffic classification because they cannot capture the complex non-linear characteristics of Internet traffic, and (2) the existing Feature Selection (FS) techniques cannot reliably provide optimal and stable features for ML algorithms. With the aim of addressing these problems, this paper presents a novel feature extraction and selection approach. First, multifractal features are extracted from traffic flows using a Wavelet Leaders Multifractal Formalism(WLMF) to depict the traffic flows; next, a Principal Component Analysis (PCA)-based FS method is applied on these multifractal features to remove the irrelevant and redundant features. Based on real traffic traces, the experimental results demonstrate significant improvement in accuracy of Support Vector Machines (SVMs) comparing to the TLS features studied in existing ML-based approaches. Furthermore, the proposed approach is suitable for real time traffic classification because of the ability of classifying traffic at the early stage of traffic transmission.

1. Introduction

Internet traffic classification has attracted much interest from a variety of technological areas [1, 2], including performance monitoring, QoS, user behavior analysis, user accounting and intrusion detection. The traditional port-based classification approach has been recognized as being inaccurate in recent years as more and more applications, such as Peer-To-Peer (P2P) and Instant Messenger (IM), have adopted dynamic communication techniques with unfixed TCP/UDP ports to overcome the performance limitations of traditional transmitting architectures [3].

As an alternative, the payload-based classification approach has a very accurate Deep Packet Inspection (DPI) mechanism, which has been widely deployed in industry [4, 5]. However, this approach not only imposes significantly higher computational complexity but also requires specific knowledge of the target application protocols in advance. Furthermore, developing signature databases becomes more expensive because of frequent revisions of the applications. Unfortunately, the problems have become even more critical as the application developers increasingly adopt data security transmission techniques, such as tunneling transferring