Computer Networks 119 (2017) 1–16

Contents lists available at ScienceDirect

Computer Networks

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

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*

ARTICLE INFO

Article history:
Received 6 July 2016
Revised 10 February 2017
Accepted 10 March 2017
Available online 11 March 2017

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.

© 2017 Elsevier B.V. All rights reserved.

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 mechanisms and traffic encryption algorithms, to prevent management devices from identifying and shaping traffic correctly.

Due to the limitations of the port-based and payload-based approaches, recent research efforts have been dedicated to developing novel classification approaches. One of the major directions is based on ML techniques [6–8], which classify traffic flows by their underlying features as derived from flow statistics. Instead of analyzing the payload data, this technique exploits only the Transport Layer Statistics (TLS) features (e.g., the packet size statistics and flow duration), which capture the distinctive usage and interaction patterns of network applications, and thus, the user privacy is guaranteed. In recent years, a variety of TLS features have been exploited for ML-based traffic classification [4,8,13,42–46]. However, these linear statistical features are not very robust for traffic classification because they cannot capture the complex non-linear characteristics of Internet traffic, which have a large impact on traffic classification. Therefore, one of the key challenges (in extracting the representative features of traffic) is to find an efficient feature extraction method that can extract the robust linear and non-linear characteristics of Internet traffic simultaneously.

The other key challenge (for traffic classification) is to search the optimal and stable features for ML algorithms. The presence of irrelevant and redundant features increases the storage require-

[*] Corresponding author.

http://dx.doi.org/10.1016/j.comnet.2017.03.011
1389-1286/© 2017 Elsevier B.V. All rights reserved.

ments and the execution time of ML algorithms and, finally, degrades their classification performance [9,10]. With the object of addressing this problem, FS techniques [14,16] can play an important role in removing irrelevant and redundant features. Despite the vast number of FS techniques proposed in the literature [14–18], obtaining the optimal features remains a challenge because of the following: (1) due to the lack of having a unified evaluation criterion, different FS techniques could select the features that may be local optima in the feature space; (2) most of the FS techniques were developed with a focus on improving the classification accuracy by removing the irrelevant and redundant features, but they neglect the stability of the feature selections for variations in the training data set.

With the aim of addressing the first challenge, multifractal analysis can play an effective role in depicting the linear and nonlinear characteristics of Internet traffic. In fact, the high variability and burst nature of Internet traffic has been reported by a number of tele-traffic research papers [19–22] for at least two decades. Due to the non-linear nature of the traffic, the traditional linear feature extraction techniques cannot work effectively. Therefore, fractal theory has been applied to describing the complex nonlinear behavior of traffic, in which Long-Range Dependence (LRD) and self-similarity have been analyzed intensively, and a group of researchers has focused on the detection of these properties and their engineering impact in the network performance and operation [23,24]. However, all of these methods reflect only the overall irregularity of the signals, without describing the local scaling properties. After a number of new measurements and in-depth analyses of Internet traffic, it has been discovered that the traffic reveals a highly irregular local structure with a more complex scaling behavior, which can be explained by multifractal features [25–27]. Multifractal features, such as multifractal spectra, can fully display the distribution of a signal's singularities, while the local scaling behaviors are depicted more precisely [28,29]. Currently, multifractal analysis has been one of the most popular tools used to depict Internet traffic.

With respect to the second challenge, a novel FS algorithm based on Principal Component Analysis (PCA) is developed to provide optimal and stable features for traffic classification. In contrast to the existing FS techniques, which lack a unified evaluation criterion, PCA is the predominant multivariate statistical technique that is aimed at reducing the dimensionality of the data, to simplify the subsequent analysis and allow for an explanation of the data in an intuitive manner [30]. The optimality properties of PCA have attracted research on PCA-based FS methods [31–33]. However, most of these methods are unsupervised because the relationship between the principal components and class labels cannot be analyzed, and they could obtain a substantial amount of redundant information because the principal components are independent of one another and the relevance between the features that explain different principal components cannot be considered. In this study, the novel FS method proposed here can find the optimal and stable features for traffic classification by exploiting the structure of the principal components with class labels.

Finally, by combining the solutions described above, this paper proposes a novel feature extraction and selection approach for traffic classification, in which multifractal features are utilized as alternatives of TLS features to depict Internet traffic, and a PCA-based FS (PCABFS) method is proposed to optimize the multifractal features. The proposed approach is evaluated using real traffic traces, and extensive experiments show that the proposed approach indeed provides efficient and robust features for traffic classification by extracting and selecting the multifractal features of Internet traffic.

The significant contributions of this paper are the following:

(1) To accurately characterize the Internet traffic, this paper proposed a mulitifractal feature extraction method that is based on wavelet leaders. This method can efficiently extract the multifractal features of traffic flows, which are more robust and discriminative than TLS features for traffic classification. Based on our extensive research, this paper presents the first attempt to use wavelet leaders multifractal features to address traffic classification issues.

(2) To overcome the problems of existing FS techniques, this paper developed a PCA-based FS (PCABFS) method for feature selection. Based on PCABFS, the impact of each feature on correct classification can be evaluated, and the relationships among the different features can be analyzed clearly. By comparison with existing FS methods, PCABFS not only has a better effect on the feature selection but also has a higher degree of stability.

(3) This paper analyzed the differences in the multifractal features between different traffic classes and provided the reasons for these differences. Furthermore, based on the observation of the multifractal features and the analysis of PCABFS, the larger Holder exponents and corresponding multifractal spectra are considered to be the optimal and robust features for traffic classification.

(4) This paper evaluated the impact of the sampling interval, sampling duration and spacing value of orders q on the classification performance. The results show that for TCP flows, traffic data with a 20 ms sampling interval and 15 s sampling duration obtained the best classification performance, while for UDP flows, traffic data with a 50 ms sampling interval and 20 s sampling duration obtained the best classification performance. In addition, the classification performance improved slightly with a decrease in the spacing value of orders q (i. e., increasing the number of multifractal features). Therefore, the sampling interval and sampling duration of the traffic flows have a large influence on the multifractal features and finally affect the classification performance.

(5) This paper compared the classification performance and runtime performance between the multifractal features obtained by the proposed approach and the TLS features studied in the previous research literature. The results show that the classification performance of the multifractal features is higher than that of the TLS features. This finding proves that multifractal features are more robust and discriminative than TLS features for traffic classification. Furthermore, although the runtime performance of the multifractal features is not higher than that of all of the TLS feature sets, the multifractal features are more effective for real time traffic classification than the TLS features due to the ability of classifying traffic at the early stage.

This paper presents an overview about related work in Section 2. In Section 3, we elaborate on the feature extraction and selection approach. Section 4 details the experimental data sets and experimental results. Finally, the study's conclusions are presented in Section 5.

2. Literature review on the relevant research

Many different techniques have been proposed for traffic classification. From the perspective of feature extraction, traffic classification approaches can be divided into two categories: traffic classification by payload inspection and traffic classification without payload inspection. In the following sections, we investigate these two types of traffic classification approaches in detail.

2.1. Traffic classification by payload inspection

This type of technique is based on inspecting the payload data in multiple network layers [34,4]. In this way, the payload data is checked in a bitwise manner to find specific bit sequence that represents the signature of a certain protocol. The protocol type of a packet can be accurately identified if such a bit sequence is found. Practically, payload-based approach is usually exploited to detect abnormal traffic [36,37] and P2P flows [3,35].

Although having particularly high accuracy, payload-based approach has many disadvantages because (1) it is a complicated operation that demands a large amount of computation and storage capacity, and usually specific hardware is necessary for inspecting the signatures in the payload data by monitoring the traffic passively or initiatively [38]; (2) it can fail to identify the protocol type if the tunnel or encryption techniques are applied to traffic transmission, and it may become completely ineffective if the new applications with new signatures are encountered; and (3) it has a very serious problem that is related to user privacy violation and laws issues.

2.2. Traffic classification without payload inspection

In contrast with payload-based approach, there are many approaches that classify traffic without inspecting the payload data. Early approaches examine the information in the network layer and transport layer, while further approaches analyze the transmission patterns of the applications.

2.2.1. Port-based approaches

Port-based approaches typically classify Internet traffic according to the 16-bit port numbers in transport layer header, which are utilized by computers to identify different communication sessions. The mappings between port numbers and services are regulated by the Internet Assigned Numbers Authority (IANA) [39]

Early researchers have relied on the simple concept that all services utilize the standard port numbers that are registered by IANA for traffic transmission [40]. However, this method proved to be inefficient because the definitive mapping is no longer effective. More and more newly emerging services violate the port number assignments of IANA. Furthermore, standard services can disguise themselves by using port numbers that are usually associated with different services, to circumvent the access control restrictions of proxies and firewalls [41,42]. Thus, port-based approaches can fail to classify the traffic if no matching port number is found or produce false results if an application uses another well-known port number. A research study conducted by Moore et al. [4,11] reported that the overall accuracy achieved by port-based approaches is only between 50% and 70%.

2.2.2. ML-based approaches

The limitation of port numbers has motivated the consideration of exploiting new TLS features to classify Internet traffic. This consideration is based on the assumption that applications typically send data in some type of pattern, which can be utilized as a communication means to differentiate network connections. It is notable that port numbers are no longer exclusively utilized to classify Internet traffic, although they are still important features for network communication. Generally, these TLS features can be obtained by directly extracting the information from network layer headers. In practice, traffic flows can be collected by built-in network devices. Because 1000 MBs of traffic lead to TLS data in the order of kBs, ML-based approaches are more efficient for traffic classification in high-speed network than payload-based approaches.

With an increasing number of TLS features to be analyzed, substantial effort has been made on the application of ML algorithms to TLS features for traffic classification [4,8,13]. Table 1 lists the different TLS feature sets that are exploited by the previous ML-based approaches. In 2004, Roughan et al. [42] and McGregor et al. [43] first extracted TLS features, e.g., the packet size statistics and flow duration, for traffic classification. In 2005, Moore et al. [4] presented 248 TLS features to fully describe the Internet traffic characteristics. In recent years, more and more TLS features were presented for characterizing network transmission behavior [8,46]. So far, hundreds of TLS features can be utilized for traffic classification.

Moreover, some research studies [1,50,8,12,51] have provided a performance evaluation of different ML algorithms in traffic classification. The evaluation results showed that SVMs, Bayesian Networks (BN), C4.5 Decision Trees (DT) and Naive Bayesian Trees (NBT) achieved the highest accuracy, but NBTs consumed longer build time compared to the other algorithms. Furthermore, several recent research studies have suggested that feature selection is more important to optimizing the performance of traffic classification than the choice of ML algorithm [47]. The implementation of FS techniques for ML-based traffic classification has become a new direction in traffic classification and has received significant attention over the past few years [48,49].

2.3. Summary and conclusions of previous studies

ML-based approach is a very promising technique for traffic classification. On the one hand, it does not rely on payload inspection, which has drawbacks of low accuracy and high computational overhead. On the other hand, it exploits many TLS features that are directly extracted from traffic, which can provide better classification results than port-based approach.

A large number of research studies [8,13,48,49] have examined the impacts of different TLS features and ML algorithms on the performance of traffic classification, and some of them found that robust TLS features are more conducive to improving the classification performance than powerful ML algorithms. Nevertheless, it can be found from Table 1 that there is no consensus on the robust TLS features in the previous studies. For example, in recent study [48], Zhang et al. choose four TLS features as the robust features from the Cambridge traffic data sets [4], for traffic classification using FS algorithms, but in another recent study [49], Fahad et al. choose five TLS features as robust features from the same traffic data sets, and most of them are different from the features chosen by Zhang et al. This finding is mainly caused by three reasons. First, many TLS features collected from traffic flows of different traffic classes are similar because they are simple linear statistics of traffic flows. Second, TLS features are not very stable to network environmental variation, especially to concept drift of Internet traffic. Finally, TLS features cannot depict the non-linear characteristics of Internet traffic, which contain some of the essential information for traffic classification. However, most recent ML-based traffic classification approaches have failed to concern the robustness problem of the TLS features in traffic classification. Moreover, it must be noted that TLS feature-based traffic classification approaches cannot perform traffic classification before transmitting all packets of a flow. In this study, we use wavelet leaders multifractal features instead of TLS features to depict the traffic flow, which not only are more robust than TLS features but also can perform classification at the early stage of transmission.

ML algorithms rely on a set of “good” features that can provide better class separability to develop accurate classification models [48]. In many research studies [43,42,44,4,8] on traffic classification, the identification of good features relies mainly on expert knowledge or empirical analysis. However, the identification pro-

Table 1
TLS features studied by the previous studies.

Ref.	TLS features
McGregor et al. [43]	Minimum packet size, maximum packet size, quartiles of packet sizes, minimum packet size as fraction of max and the first five modes; total number of bytes; the number of transitions between the transaction mode and bulk transfer mode; interarrival time (ITA); idle time in bulk mode and in transaction mode; flow duration.
Roughan et al. [42]	Mean and variance of packet sizes; root mean square of packet sizes; mean of flow duration; mean of volume; mean and variance of packets; symmetry of a connection, advertised window sizes; throughput distribution; ITA, loss rates, latencies.
Zander et al. [44]	Mean and variance of packet sizes in a forward and backward direction; mean and variance of ITA in a forward and backward direction; total number of bytes; flow duration.
Moore et al. [4]	248 features, including port number of server, port number of client; minimum ITA; media ITA; mean ITA; first quartile of ITA; minimum packet size; media packet size; mean packet size; first quartile of packet sizes; number of packets, etc.
Li et al. [46]	Port number of server and client; total number of bytes; Average segment size; median of total bytes; variance of total bytes; total numbers of RTT samples; minimum segment size; count of packets with at least 1 byte; count of all packets with a push bit.
Yuan et al. [13]	Window size statistics; Packet size statistics; total number of packets; byte ratio of sent and received packets; protocol type.
Soysal et al. [8]	Port number of server and client; total number of packets; total number of bytes; flow duration; service type; flags type; protocol type.
Zhang et al. [48]	Server port; minimum segment size (client to server); total number of bytes sent in initial window.
Fahad et al. [49]	Server port; total number of packets with PUSH bit; total number of bytes sent in initial window from server to client; total number of RTT samples.

cess is very inefficient due to excessive manual intervention, and the selected features are not very satisfactory. With the aim of selecting a subset of good features, feature selection techniques (FS) can play an effective role in reducing the dimensionality of the feature set and removing irrelevant and redundant features [49]. However, existing FS techniques, such as Correlation-based Feature Selection (CFS) [17], Information Gain (InfoGain)[14], Chi-square [15], and others conduct a search for an optimal subset using different evaluation criteria, which could make the selected subset be local optima. Furthermore, most of them fail to consider the stability of the optimal subset for variations in the flow statistical features. To improve the performance of traffic classification, this paper proposed a novel FS algorithm that is based on PCA to select the optimal multifractal features. Compared to existing FS techniques, the FS algorithm is able to not only overcome the shortcomings that existing FS algorithms lack unified evaluation criterion but also have a higher stability.

At the end of this paper, we provide a comparison of the proposed approach with the previous studies, to illustrate the advantages of the proposed approach in terms of the classification performance and computational performance.

3. Methodology

In this section, we proposed a novel feature extraction and selection approach. This approach uses WLMF to extract the multifractal features of Internet traffic, and then, the PCABFS method is applied to these multifractal features for feature selection; finally, the selected optimal multifractal features are used for traffic classification by using ML algorithms. Fig. 1 presents the schematic diagram that illustrates the proposed approach applied in on-line real-time traffic classification. It is notable that PCABFS is used only in an off-line procedure, and the results of the feature selection are stored as optimal features for traffic classification in an on-line procedure.

3.1. Wavelet leaders based multifractal features

3.1.1. Wavelet leaders

Let $\psi_0(t)$ be an elementary function with a compact time support. For $\forall \mathrm{k} = 0,\ 1,\ldots,\ N_\psi - 1$, where the positive integer $N_\psi \geq 1$, if $\psi_0(t)$ satisfies $\int_R t^k \psi_0(t)dt \equiv 0$, and $\int_R t^{N_\psi} \psi_0(t)dt \neq 0$, $\psi_0(t)$ can be characterized by its vanishing moments N_ψ. At the same time, the collection of dilated and translated templates of $\psi_0(t)$, $\{\psi_{j,k}(t) = 2^{-j/2}\psi_0(2^{-j}t - \mathrm{k}),\ j \in Z,\ k \in Z\}$ form an orthonormal basis of $L^2(\mathrm{R})$.

For the signal $X = \{x_k,\ k \in Z)$, its discrete wavelet transform is defined by the following coefficients [56]:

$$d_x(j,k) = \int_R X(t)2^{-j}\psi_0(2^{-j}t - k)dt.$$

Gen that $\psi_0(t)$ has a compact time support, we define the dyadic interval as $\lambda = \lambda_{j,k} = [k2^j, (k+1)2^j]$, and we let 3λ denote the union of the interval λ with its two adjacent dyadic intervals: $3\lambda_{j,k} = \lambda_{j,k-1} \cup \lambda_{j,k} \cup \lambda_{j,k+1}$. Then, the wavelet leader is defined as the local supremum of the wavelet coefficients taken within a spatial neighborhood over all finer scales [55,57]: $L_X(j,k) \equiv L_\lambda = \sup_{\lambda' \subset 3\lambda} |d_{X,\lambda'}|$. Hence, $L_X(j,\ k)$ consists of the largest wavelet coefficient $L_X(j',\ k')$, which is calculated at all finer scales $2^{j'} \leq 2^j$ within a narrow time neighborhood $(\mathrm{k} - 1)2^j \leq .\ 2^{j'}k' < (k+2)2^j$.

3.1.2. Features extraction based on wavelet leaders

For signal X, let $S_L(q,\ j)$ and $\zeta_L(q)$ respectively denote the structure function and the corresponding scaling exponents, where j is the analysis scales, and q is the moment order; they can be defined as follows, respectively:

$$S_L(q,j) = \frac{1}{n_j}\sum_{k=1}^{n_j} |L_X(j,k)|^q$$

$$\zeta_L(q) = \lim_{2^j \to 0} \inf\left(\tfrac{\log_2 S_L(q,j)}{j}\right)$$

Then, the multifractal spectrum $D(h)$ of the signal X can be obtained by the Legendre transform of the scaling exponents $\zeta_L(q)$

$$D(h) = \inf_{q \neq 0}(1 + qh - \zeta_L(q))$$

To avoid the complex Legendre transform, $\zeta_L(q)$ can be estimated by method of linear regressions via $\log_2 2^j = j$ versus $\log_2 S^L(j,\ q)$ and $\ln 2^j$versus $C^L(j,\ p)$

$$\zeta_L(q) = \sum_{j=j_1}^{j_2} w_j \log_2 S^L(j,q)$$

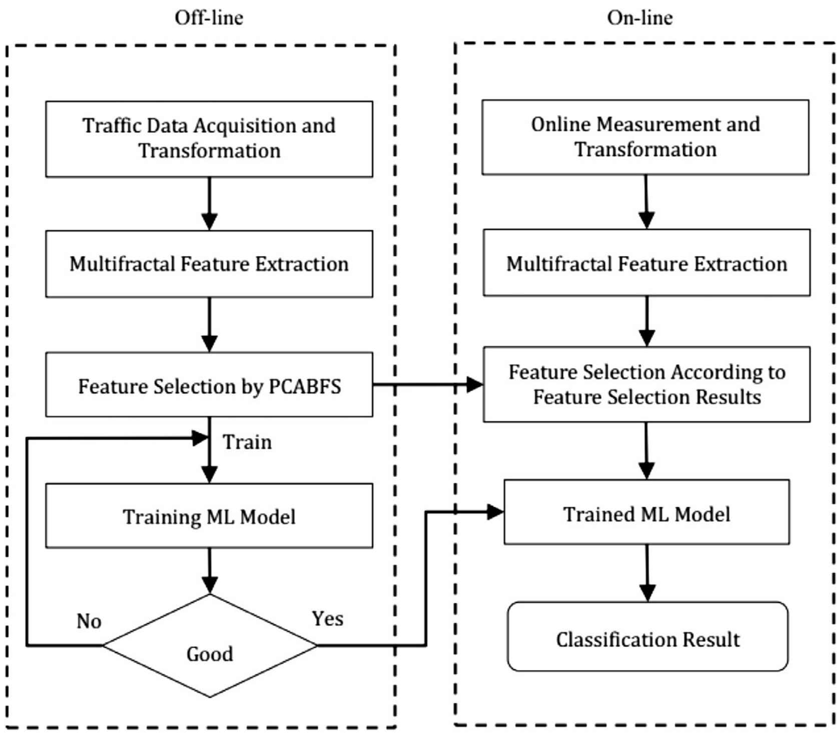

Off-line
On-line
Traffic Data Acquisition and Transformation
Multifractal Feature Extraction
Feature Selection by PCABFS
Train
Training ML Model
No
Yes
Good
Online Measurement and Transformation
Multifractal Feature Extraction
Feature Selection According to Feature Selection Results
Trained ML Model
Classification Result

Fig. 1. Flow chart of the traffic classification system.

In these formulas, a parametric formulation of Chhabra is utilized to estimate the multifractal spectrum to reduce the computational complexities [58].

In this study, the multifractal features consist of $D(h(q))$ and $h(q)$.

3.1.3. Advantages of WLMF

The WLMF method is very reminiscent of the Wavelet Transform Modulus Maxima (WTMM) technique [7], which is highly effective for characterizing real systems from statistical data and has been extensively used in many research fields [52-54]. However, according to the literatures [55,59], WLFM has some significant advantages over WTMM. First, from a mathematical point of view, WLMF has solid supports of well-established mathematical theory and the scaling exponents generated by WLMF are more precise. In respect of WTMM, due to the irregularity of the spacing between the local maxima, the scaling exponents generated by WTMM are very different from those generated by WLMF. Furthermore, there is no mathematical result to provide support for this method even now. Second, from a more practical point of view, WTMM utilizes continuous wavelet transform for finding and chaining the local maxima, which results in a very high computational cost, especially for high-dimensional signals such as Internet traffic. By contrast, WLMF utilizes discrete wavelet transform and fast pyramidal algorithm to generate multifractal features, which can significantly reduce the computational overhead for high dimensional signals. Therefore, WLMF is more suitable for Internet traffic classification than the WTMM.

3.2. Principal Component Analysis-based feature selection

3.2.1. Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate statistical method that aims at reducing the dimensionality of the data space while accounting for as much variation in the data as possible [30]. This technique is particularly useful when the variables in the data are highly correlated. The idea behind PCA is to compute new variables called principal components to achieve these goals. These variables are obtained as linear combinations of the original variables. The values of these new variables are called eigenvalues and are interpreted geometrically as the projections of the observations onto the principal components.

Singular Value Decomposition (SVD) is a linear algebra method and can be used to compute these variables efficiently [60]. Through SVD, any matrix X with n rows and m columns can be decomposed as a product:

$$X = USV^{T}$$

where U is an $n \times m$ matrix with orthonormal columns ($U^{T}U=I_{m}$), while V is an $m \times m$ orthonormal matrix ($V^{T}V=I_{m}$), and S is a $m \times m$ diagonal matrix with positive or zero elements, called the singular values. The covariance matrix of matrix X can be rewritten as follows:

$$C = \frac{1}{n}XX^{T} = \frac{1}{n}US^{2}U^{T}$$

Thus, the transformed data can be written as

$$Y = \tilde{U}^{T}USV^{T}$$

where $\tilde{U}^{T}U$ is a simple $n \times$ m identity matrix.

In this study, in addition to the multifractal features of Internet traffic, the class label is also used as an important feature (i.e., class feature) for PCA to analyze the correlation between the multifractal features and the traffic class. The value of the class label for each class is set to the product of the average value of all of the multifractal features of flows that belong to the class and a random integer, where the average value represents the overall characteristics of the class and the random integer is used to increase the differences from other classes.

3.2.2. Component selection

The component selection procedure is performed to find a component subspace of size q, which reserves most of the variance of original data set. In this procedure, the components are sorted in descending order according to their eigenvalues, and the variance ratio represented by each component and the corresponding cumulative variance ratio is respectively calculated. Then, the set of q components ($q \le m$) is selected according to their associated eigenvalues.

Recently, several methods have been used to select the optimal number of components to represent the variance of the original data. One of the most popular methods is the Kaiser Criterion [61]. This method suggests that the components with eigenvalues equal to or greater than 1 should be selected, i.e. each of the components should represent at least $1/m$ of the total variance. The variance ratio represented by a component c_i can be defined as follows:

$$v(c_i) = \frac{ev_i}{\sum_{j=0}^{n} ev_j}$$

As a result, we apply this method to the principal components to select the first q components as the subspace of the original features.

3.2.3. Orthogonal rotation

To better represent the correlation between the components and features after selecting the optimal components, the subsequent analysis usually involves a rotation of the selected components. There are two types of rotation methods are extensively used for PCA: orthogonal method (i.e. the axes after rotation are still orthogonal to one another) and oblique method (i.e. the axes after rotation need not be orthogonal). It is notable that all rotation methods do not change the inertia explained by the total subspace of components (only the partition of the inertia will be changed). In this study, VARIMAX rotation [62] is exploited to transform the original data. This method relies on the simple solution that each component has a small number of large loadings and a large number of zero (or small) loadings. Formally, VARIMAX searches for a linear combination of the original components so that the variance of the squared loadings is maximized, which amounts to maximizing v:

$$v = \sum (q_{j,l}^2 - \bar{q}_{j,l}^2)^2$$

where $q_{j,l}^2$ is the squared loading of the j'th variable of the matrix Q on the component l, and $\bar{q}_l^2$ is the mean of the squared loadings.

This rotation is performed by using the eigenvectors associated with the correlated matrix of X and the diagonal matrix of singular values.

3.2.4. Optimal feature selection

Through rotation, each component can be mainly explained by a group of multifractal features, and the loadings of the class feature represent the impact of each component on the traffic classification. In this way, the correlations of the multifractal features with the class feature become clearer. However, the optimal feature selection is still a challenge because the components are independent of one another and the relevance between the multifractal features that explain different components cannot be identified by only analyzing the components. Therefore, cluster analysis is performed on the loadings of all features to determine the optimal features.

Let vector V_i ($i \le q$) consists of the loadings of the i'*th* features in the q components. The vector V_i represents the projection of the i'*th* feature of the original matrix X to the lower dimensional space, and the q elements of V_i correspond to the weights of the i'*th* feature on each axis of the subspace. Cluster the vectors into two groups using the K-Means algorithm with Euclidean distance. Each group is composed of the features that have high relevance with respect to one another; then, the group to which the vector of class feature belongs is the optimal subset of features, and the other group is the subset of redundant features.

Table 2
Data sets statistics.

Data set	Sampling interval	Sampling duration	Flow size
D_10_20	10 ms	20 s	2000
D_10_15	10 ms	15 s	1500
D_10_10	10 ms	10 s	1000
D_20_20	20 ms	20 s	1000
D_20_15	20 ms	15 s	750
D_20_10	20 ms	10 s	500
D_50_20	50 ms	20 s	400
D_50_15	50 ms	15 s	300
D_50_10	50 ms	10 s	200

4. Experimental evaluation

4.1. Traffic data

To analyze the multifractal features of different Internet traffic in detail, two traffic traces are used for all of the experiments in this study. The first traffic trace is captured at the edge link from Qingdao Agricultural University (QAU) from 8:30 a.m. to 5:30 p.m., October 14, 2011, and the second traffic trace is captured from a PC room, which has approximately 20 computers, of the School of Information Science and Engineering, Ocean University of China (OUC), from 9:30 a.m. to 5:00 p.m., November 18, 2014. In these traffic traces, the basic data instances were IP flows, which consisted of sequences of packets that were exchanged between pairs of specific endpoints for the purpose of inter-process communication across the Internet. Thus, each flow was a time series of packets that shared the same five-tuples, which consisted of the source and destination IP addresses, the source and destination port numbers, and the transport layer protocol. In this study, we used bidirectional flows as the basic experimental object, which are a pair of unidirectional flows (i.e., upload flow and download flow) that travel in opposite directions between the same source and destination IP addresses and ports.

To evaluate the classification performance of the proposed approach, all of the flows were pre-classified by performing payload inspection to obtain the corresponding protocol label. The signature rules consisted of all of the previous experiences and some well-known tools such as L7-filter, Tstat and NeTraMark. In this study, we mainly analyzed the seven most common TCP flow classes: WWW, P2P, IM, HTTP+Flash, SMTP, POP and IMAP, and the three most common UDP traffic classes: P2P, IM and VoIP. We randomly selected 10,000 samples of bidirectional flows per traffic class from the traffic traces and transformed them to the new time series of transferred byte numbers in a fixed sampling interval. To evaluate the impacts of the sampling interval and sampling duration on the performance of the proposed approach, we created nine different data sets of the newly transformed time series according to various sampling intervals and sampling durations to analyze and compare their classification performances. These data sets statistics are presented in Table 2. Each data set has 10 × 10, 000 pairs of time series of traffic flows.

In this study, data set D_20_15 was first used to demonstrate the performance of the proposed approach, and the flows of TCP and UDP were separated to evaluate the classification performance of the proposed approach. Fig. 2 shows a typical data sample of bidirectional flows for each traffic class. Here, we should note that because the duration of the flows of POP and IMAP were shorter than 15 s, we added a zero value at the end of these time series to extend these flows to the length of 15 s. As Fig. 2 shows, it is not easy to identify the different traffic classes because some of the flows of different traffic classes, such as the flows of WWW, P2P, IM and HTTP+FLASH, are very similar.

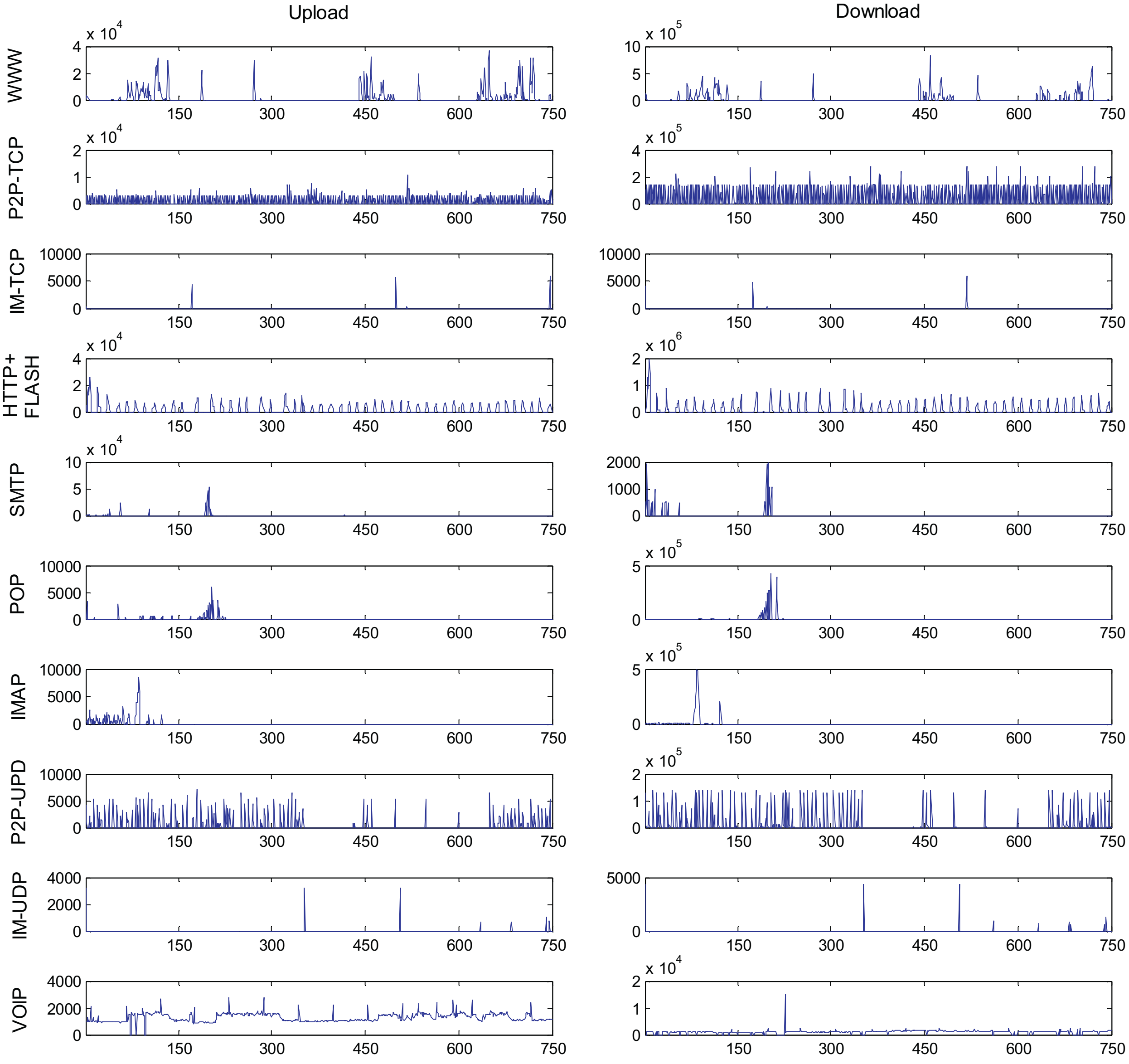


Fig. 2. Typical data samples in data set D_20_15 for 10 classes. Vertical axis: number of bytes; horizontal axis: sampling interval (20 ms).

4.2. Feature extracting using WLMF

4.2.1. Mother wavelet selection

In the WLFM method, the choice of the mother wavelet is a crucial factor that influences the precision of multifractal features that are extracted from the Internet traffic. Currently, there are many different mother wavelets that can be chosen for multifractal analysis [63]. Wendt et al. [64,65] presented that only when the vanishing moment is larger than the largest Holder exponent of the signals, the mother wavelet can stabilize the estimates of the structure functions and suppress the potentially superimposed smooth trends for the negative orders q. However, overlarge vanishing moment will result in border effects that can degrade the performance of the multifractal analysis. Therefore, a reasonable vanishing moment can be obtained when the multifractal features do not significantly change with the increase of vanishing moment. For data set D_20_15, the Daubechies wavelet of order 3 is used as the mother wavelet, which can offer an appropriate trade-off between the preservation of spatial characterization and the relief of frequency aliasing.

4.2.2. Range of orders q

To obtain the entire multifractal spectrum (for both positive and negative values of q) by performing the Legendre transform on scaling exponents. The range of orders q should be chosen to avoid the linearization effect in the transform procedure [64]. Let $q \in [-q^*, +q^*]$ $(q \neq 0)$, if for both $q \geq +q^*$ and $q \leq -q^*$, $\zeta_L(q)$ is a linear function of q, the range $[-q^*, +q^*]$ is the reasonable range of orders q, and the approximation of q^* can be calculated by $\sqrt{2/|c_2|}$ [65]. In this study, we found that the flows of different traffic classes had different c_2 by performing calculations. To guarantee the range $[-q^*, +q^*]$, we can obtain all of the multifractal features of all of the flows and set a unified q^* for the whole data set D_20_15. To simplify the calculation process of the proposed approach, c_2 was estimated on the P2P TCP flows in data set D_20_15 by bootstrap technology, which had the minimum absolute value. As shown in Fig. 3, the minimum c_2 of the P2P TCP flows was approximately 0.086, and as a result, we allowed the range of q to be $[-5, +5]$.

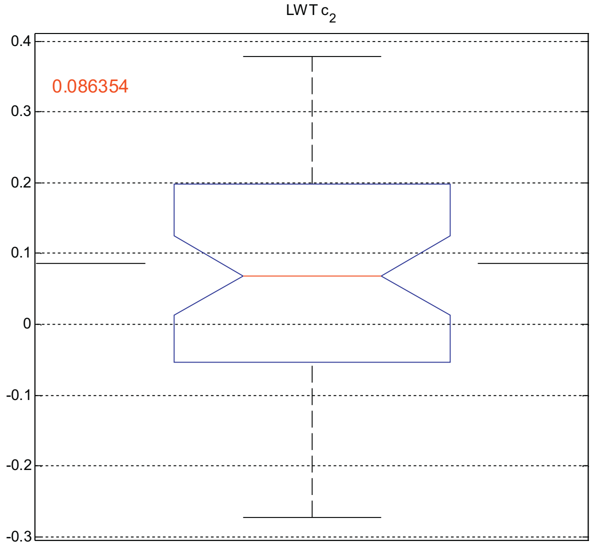


Fig. 3. Estimation of c_2 on the P2P TCP flows in D_20_15.

4.2.3. Scaling range of regression

To avoid the complex Legendre transform, the linear regression method was used with $S_L(j, q)$ over a certain scaling range $[j_1, j_2]$ to estimate *D(h(q))* and $h(q)$. Additionally, the scaling range can be estimated by statistical procedures [65]. In this study, the plot of $\zeta_L(q)$ versus $\log_2 2^j = j$ decides the scaling range in which the scaling behavior of the data can be full described. Certainly, the scaling range should be the same for all scaling exponents that are extracted from original data sets. Taking the data set D_20_15, for example, the scaling range was from 3 to 6.

4.2.4. Feature extraction

Due to the differences in the transmission mechanism of various Internet applications, the multifractal features of their traffic, such as singularity distributions and local scaling behaviors, would be different from one another. To analyze the multifractal features of different Internet traffic, data samples of different traffic classes in the data set D_20_15 were conducted by WLMF. In addition, to facilitate the calculation of multifractal features, the spacing value of the orders q was set to 1.0, and the number of multifractal features (i.e., $D(h(q))$ and $h(q)$) of each flow was 20. Fig. 4 shows the multifractal features of the data sample for each traffic class, and the corresponding traffic flows are shown in Fig. 2. From Fig. 4, it can be found that similar to their traffic flows, the multifractal features are also difficult to be classified. However, we found that the multifractal features of the WWW, IMTCP and IMUDP flows had larger Holder exponents, the largest Holder exponents of which were more than 6.5. The reason for these results was that the traffic of WWW, IMTCP and IMUDP was mainly generated by user activities, such as clicking over web page links or sending an instant message, which were related to actions of larger time scales and resulted in a higher bursty nature of the traffic. Nevertheless, the multifractal features of the other class traffic flows had smaller Holder exponents, which were less than 2.5. This finding occurred because this traffic was mainly generated by the transmission mechanism of the protocols, such as the session, queuing and transmission control, which was related to actions of small time scales and resulted in a lower bursty nature of the traffic. Therefore, by observation, we considered that the higher fractal dimensions and the corresponding multifractal spectra, namely, $h(q)$ and $D(h(q)$ $(q < 0)$, in the multifractal features were more conducive for traffic classification. In the next section, PCABFS is utilized to search for the optimal multifractal features.

Table 3
First 10 components from the PCA.

Component	Eigenvalues	Variance (%)	Cumulative variance (%)
1	14.99	36.57	36.57
2	12.74	31.07	67.64
3	7.00	17.09	84.73
4	2.74	6.69	91.42
5	1.44	3.51	94.93
6	0.60	1.46	96.39
7	0.51	1.24	97.64
8	0.37	0.90	98.54
9	0.17	0.43	98.97
10	0.14	0.35	99.31

4.3. Feature selection using PCA

4.3.1. PCA of multifractal features

In this section, we utilized the PCABFS to analyze and optimize the multifractal features of the Internet flows. To facilitate this calculation, we cascaded the multifractal features of the upload and download flows. We first subjected the multifractal feature values of seven TCP traffic classes in data set D_20_15 and their class feature values (a total of 41 feature values for each sample) to the PCA.

Table 3 presents the first 10 components and their associated eigenvalues, variance ratio and cumulative variance ratio. In the original space, each of the 41 features contained approximately 2.44% (1/41) of the total variance. The principal components that are selected should explain at least 2.44% of the total variance. Therefore, the first five components, which accounted for 94.93% of the total variance, were selected to model the data. Next, we subjected the multifractal feature values of three UDP traffic classes in data set D_20_15 and their class features values to the PCA, and the first 4 components, which accounted for 94.23% of the total variance, were selected to model the data.

4.3.2. Feature selection by relevance

To further clarify the relationship between the principal components and multifractal features, an orthogonal rotation of VARIMAX was performed on the selected components. Thus, each component can be mainly explained by a group of multifractal features, and the loading of the class feature in each component represents the projection on each axis of the subspace, which indicates the impact of each component on the traffic classification.

We clustered the vectors that consisted of the loadings of each feature in the five components into two groups using the K-means algorithm, and the result shows that the multifractal features of $D(h(q))$ $(-5 \le q \le -1)$ and $h(q)(-5 \le q \le -1)$ of the upload flows and download flows and the class feature were clustered into a group, and the other multifractal features were clustered into the other group. Therefore, the multifractal features of $D(h(q))$ $(-5 \le q \le -1)$ and $h(q)(-5 \le q \le -1)$ of the upload flows and download flows were the optimal features of TCP the traffic flows in data set D_20_15. In other words, for the TCP flows in data set D_20_15, the larger Holder exponents and the corresponding multifractal spectra of the upload flows and download flows can better represent the natures of the traffic flows, and the other multifractal features are the redundant features, which should be removed. These results are consistent with the observation results in Section 4.2.4. To intuitively illustrate the relationship between the optimal multifractal features and the components, Fig. 5 shows the component mapping between component 1 and component 2. As Fig. 5 shows, $D(h(q))$ $(-5 \le q \le -1)$ and $h(q)(-5 \le q \le -1)$ of the upload flow and download flow respectively explain most of the variability of Components 1 and 2, and all of them have a high correlation with the class features.

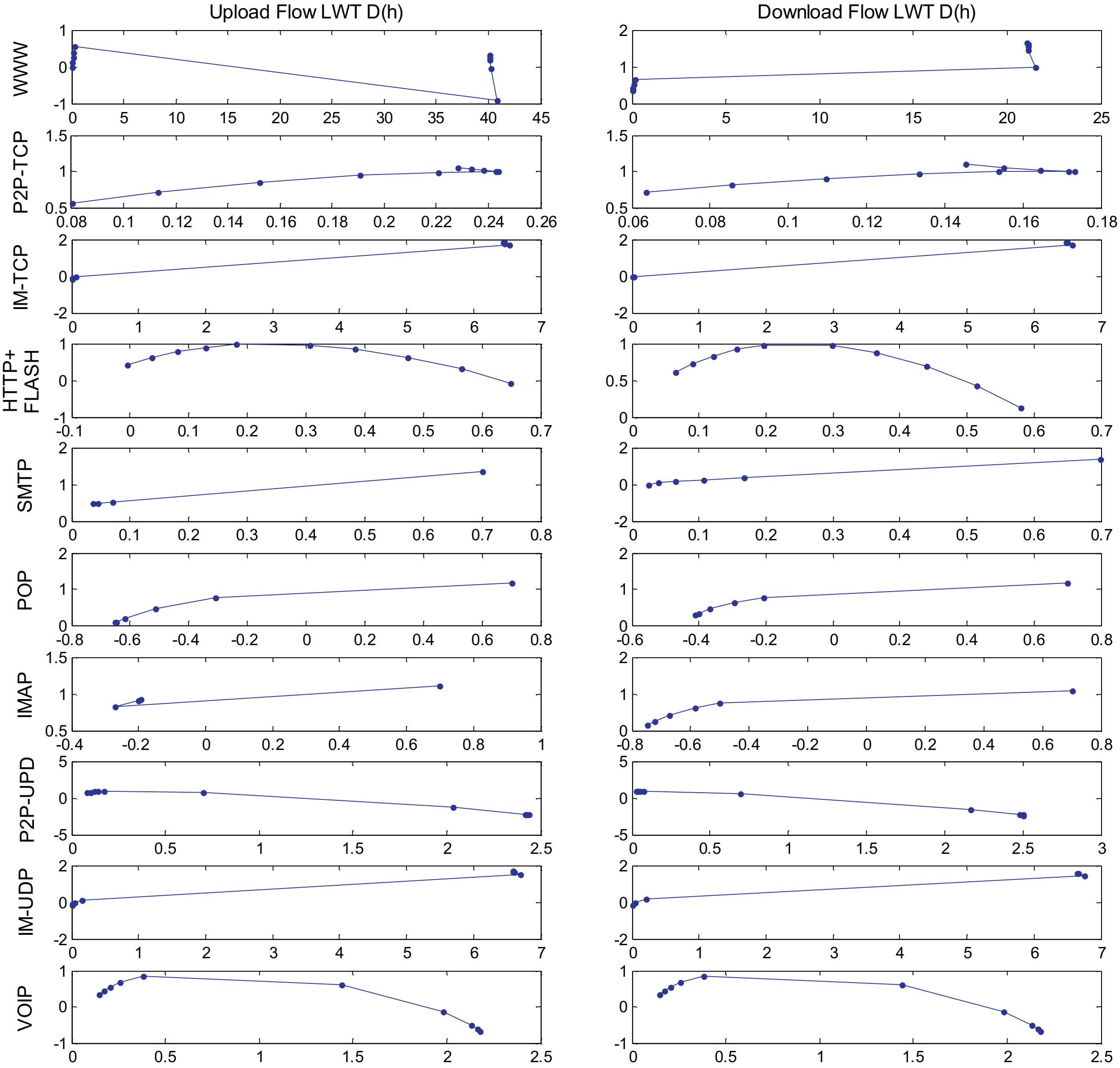


Fig. 4. Multifractal features for 10 different classes of traffic flows. Vertical axis: $D(h(q))$; horizontal axis: $h(q)$.

Then, by using the same method, the same result was obtained for the UDP flows in data set D_20_15.

4.4. Classification performance

4.4.1. Performance evaluation of the proposed method

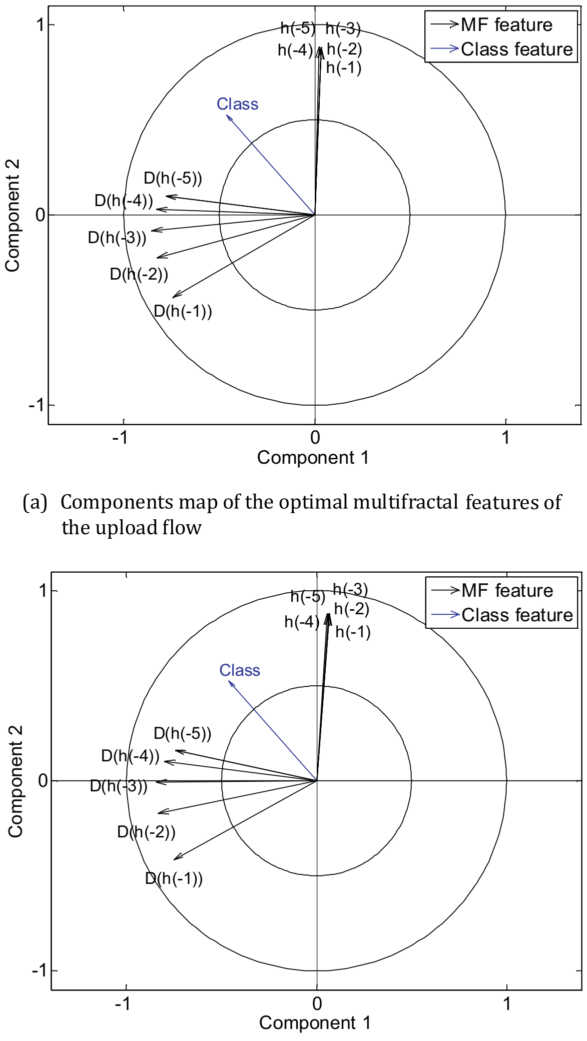
In this section, we concentrate on the classification performance of the proposed approach and analyze the impact of the sampling interval and sampling duration. The SVMs was chosen to evaluate the classification performance, as it benefits from well-established principles of structural risk minimization and has the advantage of favorable generalization ability in the field of intelligent classification [66,67]. For each data set in Table 2, the classification accuracy of the proposed approach is calculated by SVMs over a 10-fold Cross-Validation (CV) [68]. Table 4 gives the classification accuracy of the proposed approach for nine data sets. As can be observed, there are many differences in the classification accuracy of the nine data sets, and therefore, the sampling interval and sampling duration of the traffic flows have a large influence on the classification performance.

When comparing the classification results for different data sets, for TCP flows, the proposed approach obtained the highest overall accuracy for data set D_20_15, which is 95.67%. However, there were still approximately 10.8% WWW flows and approximately 12.3% IM flows that were misclassified, which degraded the overall accuracy. For most of the other data sets, the misclassifications of the WWW, IM, P2P, and HTTP+Flash flows were the main reasons that led to the lower overall accuracy, and only for the data sets D_20_20 and D_50_20, the misclassifications of the SMTP and POP flows were very serious. In addition, for the UDP flows, the proposed approach obtained a higher overall accuracy for the data sets D_20_20, D_20_15 and D_50_20, which had more than 95.0% accuracy. For most of the other data sets, the misclassifications of the P2P and IM flows were the main reasons that led to the lower overall accuracy, but for the three data sets with10 ms sampling intervals, the misclassifications of all of the three classes of traffic flows were more serious.

Moreover, for the TCP flows, there was not a clear trend between the classification accuracy and the sampling intervals and sampling duration. However, we found that corresponding to the

Table 4
Classification accuracy of the proposed approach for nine data sets.

Data set	Accuracy rate (%)											
	TCP flows								UDP flows			
	WWW	P2P	IM	HTTP+ FLASH	SMTP	POP	IMAP	Mean	P2P	IM	VOIP	Mean
D_10_20	59.7	99.5	84.7	99.8	94.3	99.2	98.3	90.79	88.6	86.6	92.0	89.07
D_10_15	76.0	96.7	91.3	98.0	91.7	99.2	98.8	93.10	86.3	94.6	93.4	91.43
D_10_10	72.7	92.7	89.0	91.3	98.8	99.2	99.5	91.89	91.4	91.1	85.4	89.30
D_20_20	69.0	92.2	74.3	99.3	0.0	23.0	92.7	64.30	99.9	92.6	99.6	97.37
D_20_15	89.2	98.7	87.7	99.3	96.2	99.8	98.8	**95.67**	99.9	91.1	99.9	96.97
D_20_10	86.2	89.7	85.2	87.8	99.3	99.2	98.0	92.20	93.7	83.7	99.9	92.43
D_50_20	93.3	68.7	47.2	99.8	62.8	97.8	98.7	81.19	93.1	99.9	100.0	**97.67**
D_50_15	82.7	60.7	84.0	99.0	92.7	91.2	96.8	86.73	84.0	99.9	99.1	94.33
D_50_10	69.3	48.3	72.7	55.7	98.7	99.3	98.7	77.53	80.6	99.9	95.4	91.97



(a) Components map of the optimal multifractal features of the upload flow

(b) Components map of the optimal multifractal features of the download flow

Fig. 5. Components map of the optimal multifractal features.

same sampling interval, the overall accuracy for the data sets with 15 m sampling duration were higher than that for the data sets with other sampling durations, and corresponding to the same sampling duration, the overall accuracy for the data sets with 20 ms sampling interval were higher than that for the data sets with other sampling intervals, except for data set D_20_20. Nevertheless, for the UDP flows, there was a clear trend in that the overall accuracy increases with increases in the sampling interval and sampling duration of the data sets.

In fact, decreasing the sampling interval of the traffic flows can increase their temporal resolution and thereby enhance their bursty nature, which leads to an increase in the Holder exponents of the traffic flows, and vice versa [19,21,26,27]. Therefore, the traffic flows that have a too-small sampling interval, such as the WWW and P2P UDP flows in the data sets D_10_20, D_10_15 and D_10_10, or a too-large sampling interval, such as the P2P TCP and P2P UDP flows in the data sets D_50_20, D_50_15 and D_50_10, could make their multifractal features similar to those of other traffic flows. In addition, shortening and extending the sampling durations of the traffic flows can change their shapes and thereby change their multifractal features. Therefore, the traffic flows that have a too-short sampling duration, such as the HTTP+Flash flows in the data sets D_10_10, D_20_10 and D_50_10, or a too-long sampling duration, such as the IM TCP and SMTP flows in the data sets D_20_20 and D_50_20, could make their multifractal features similar to those of the other traffic flows. Only traffic flows that have an appropriate sampling interval and sampling duration, such as all of the TCP flows in data set D_20_15 and all of the UDP flows in data set D_50_20, can show the unique multifractal features for different traffic classes.

4.4.2. Impact of the number of multifractal features

To analyze the impact of different numbers of multifractal features, which were obtained by the proposed approach, on the classification performance, we increased the number of multifractal features by reducing the spacing value of orders q to 0.5 and 0.25, which was set to 1.0 previously. For ease of interpretation, we limited the discussion to only the representative cases of the TCP flows in data set D_20_15 and the UDP flows in data set D_50_20. Tables 5 and 6 show the classification accuracy of different numbers of multifractal features for the TCP flows and UDP flows, respectively.

As Tables 5 and 6 shown, for both the TCP flows and UDP flows, the overall accuracy had a slight improvement when the number of multifractal features increased, and the highest overall accuracy was obtained when the number of multifractal features was 80. This implied that as the number of multifractal features increased, more detailed features can be obtained for improving classification performance.

4.4.3. Performance of the original multifractal features

To demonstrate the effectiveness of the PCABFS method, the original multifractal features, without applying PCABFS, were conducted by SVMs to obtain the classification accuracy. Tables 7 and 8 show the classification accuracy of different numbers of the original multifractal features that were extracted from the TCP flows in data set D_20_15 and the UDP flows in data set D_50_20, respectively. As the results shown in Tables 7 and 8, for both the

Table 5
Brief summary of the overall accuracy of different numbers of the multifractal features for the TCP flows.

Data set	Spacing value of orders q	Number of optimal features	Accuracy rate (%)			
			Max	Min	Median	Mean
D_20_15	1.0	20	97.71	91.57	95.07	95.67
	0.5	40	97.86	91.86	95.39	95.96
	0.25	80	97.57	91.71	95.11	95.77

Table 6
Brief summary of the overall accuracy of the different numbers of the multifractal features for the UDP flows.

Data set	Spacing value of orders q	Number of optimal features	Accuracy rate (%)			
			Max	Min	Median	Mean
D_50_20	1.0	20	100.0	93.33	99.83	97.67
	0.5	40	100.0	94.0	100.0	97.81
	0.25	80	100.0	95.33	100.0	97.95

Table 7
Brief summary of overall accuracy of different numbers of the original multifractal features for the TCP flows.

Data set	Spacing value of orders q	Number of multifractal features	Accuracy rate (%)			
			Max	Min	Median	Mean
D_20_15	1.0	40	96.0	78.14	88.71	88.31
	0.5	80	96.57	76.57	86.57	87.63
	0.25	160	96.57	76.29	86.00	86.83

Table 8
Brief summary of overall accuracy of different numbers of the original multifractal features for the UDP flows.

Data set	Spacing value of orders q	Number of multifractal features	Accuracy rate (%)			
			Max	Min	Median	Mean
D_50_20	1.0	40	100.00	75.67	92.67	92.15
	0.5	80	100.00	75.67	89.67	89.87
	0.25	160	98.67	74.67	87.33	87.89

TCP and UDP flows, the overall accuracy of the original multifractal features was lower than that of the features that were selected by the PCABSF method, and the overall accuracy decreased when the number of the original multifractal features increased.

In fact, although the original multifractal features can show the differences in the transmission mechanism between various Internet applications, these features contain some redundant features, such as the smaller Holder exponents and the corresponding multifractal spectra, which could seriously degrade the classification accuracy. Therefore, the PCABSF method proposed in this study can play an effective role in improving classification accuracy.

4.5. Performance evaluation of PCABFS

In this section, we provide a performance comparison of PCABFS to the existing FS methods, which include InfoGain [14], Chi-Square [15], Fisher-Score [16], CFS [17] and Fast Correlation-Basd Filter (FCBF) [18], to evaluate the performance of PCABFS. Currently, there are limited comparative research studies about FS methods in the literature [69,70]. In general, when evaluating the performance of FS methods, two aspects are of concern: (i) the accuracy (i.e., how well a generated subset can classify the traffic flows accurately), and (ii) the stability (i.e., the property of selecting the same set of features irrespective of the variations in the subset of data taken (for training) from the whole data set) [70]. Therefore, we evaluated the performance of PCABFS according to these two aspects.

4.5.1. Classification performance evaluation

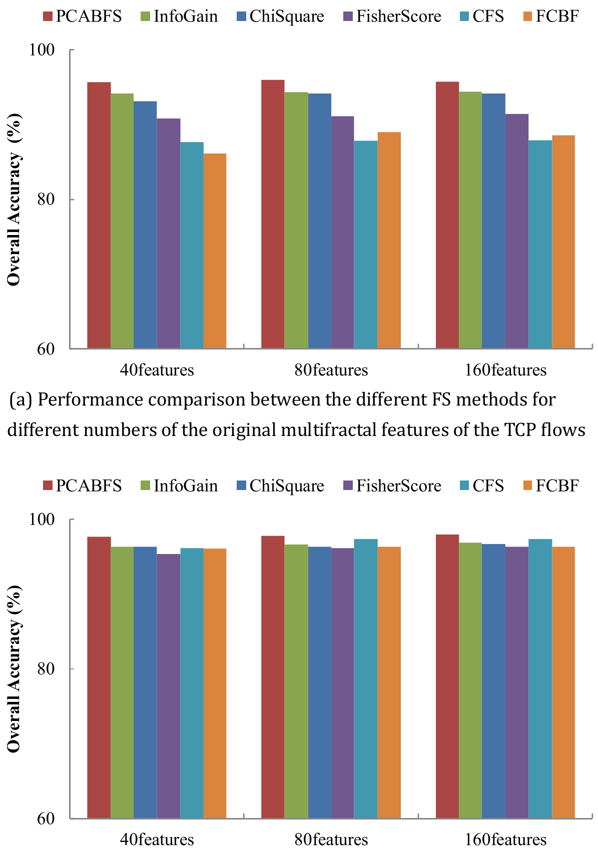
In this experiment, all of the FS methods were applied on different numbers of original multifractal features that were extracted from the TCP flows in data set D_20_15 and the UDP flows in data set D_50_20, and the overall accuracy of each FS method were calculated over a 10-fold CV.

Fig. 6 shows the comparison of the overall accuracy between different FS methods. From Fig. 6a, it can be found that PCABFS obtained the best overall accuracy for all three conditions for TCP flows. However, CFS and FCBF obtained the lower overall accuracy compared to other FS methods. Furthermore, from Fig. 6b, we found that all of the FS methods had very high overall accuracy for all three conditions for UDP flows, and the overall accuracy of PCABFS was slightly higher than that of other FS methods.

4.5.2. Stability evaluation

In this part, we evaluate the stability of all candidate FS methods. The stability of FS algorithms was defined by Kalousis [70] as the robustness of the feature preferences that are generated in response to differences in the training sets that are drawn from the same generating distribution. A high stability of an FS is crucial for reliable results and equally important for high classification accuracy when evaluating the FS performance [70,71]. In this study, the stability measure introduced in [49] was used to evaluate the different FS methods.

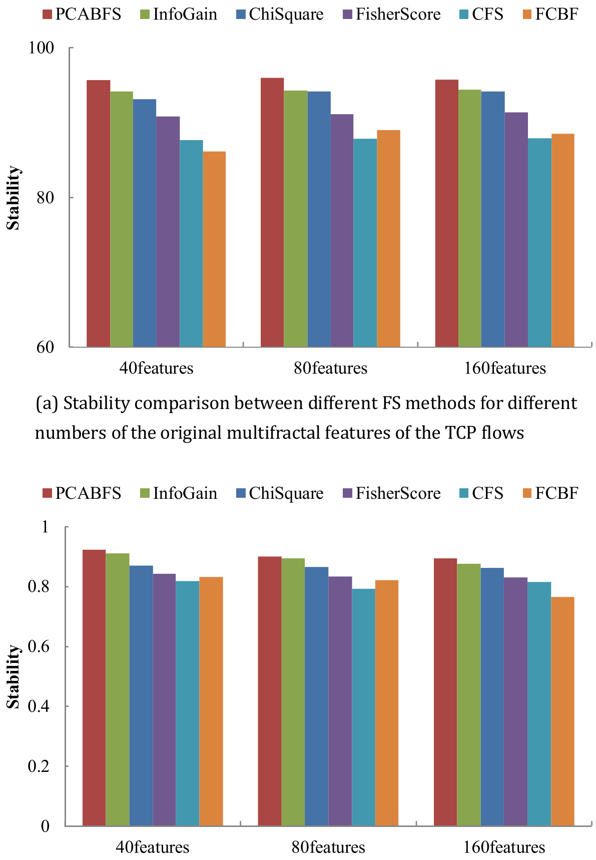
Fig. 7 shows the stability results that were obtained by different FS methods for different numbers of original multifractal features that were extracted from the TCP flows in data set D_20_15 and the UDP flows in the data set D_50_20. As Fig. 7 shows, for both the TCP and UDP flows, PCABFS obtained the highest stability for all three conditions. Furthermore, InfoGain and Chisquare obtained the second and third highest stabilities, respectively, which are slightly lower than PCABFS. Other FS techniques obtained lower stabilities for all three conditions.



(a) Performance comparison between the different FS methods for different numbers of the original multifractal features of the TCP flows

(b) Performance comparison between the different FS methods for different numbers of the original multifractal features of the UDP flows

Fig. 6. Performance comparison between different FS methods for different numbers of the original multifractal features.



(a) Stability comparison between different FS methods for different numbers of the original multifractal features of the TCP flows

(b) Stability comparison between different FS methods for different numbers of the original multifractal features of the UDPflows

Fig. 7. Stability comparison between different FS methods for different numbers of the original multifractal features.

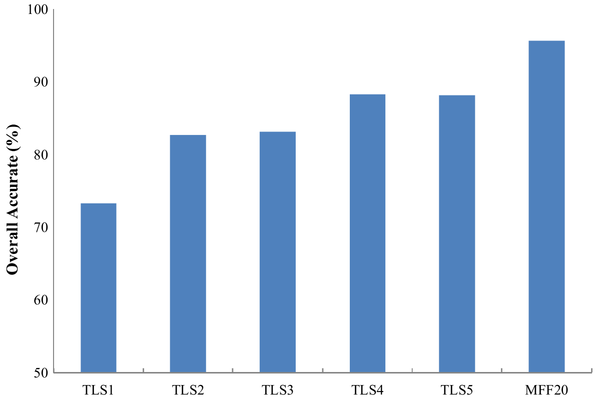


Fig. 8. Classification performance comparison between the TLS features and the multifractal features.

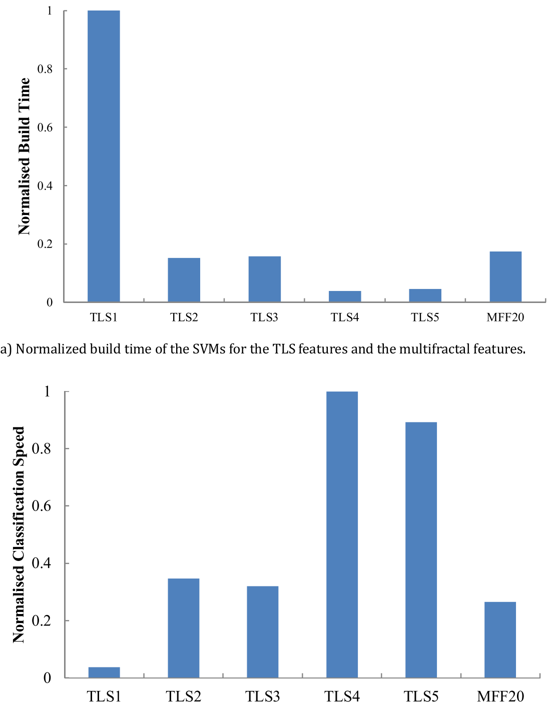
Ultimately, by comparing Figs. 6 and 7, we believe that PCAFS has the best performance for traffic classification because it obtains the highest overall accuracy and stability for all of the conditions.

4.6. *Performance Comparison of the Proposed Approach to Previous Work*

To further prove the effectiveness of the proposed approach, this section provides a performance comparison between the multifractal features obtained by the proposed approach and the TLS features studied in the previous literature, which was reviewed in Section 3.2.2, with respect to the classification performance and runtime performance.

4.6.1. Classification performance comparison

In this section, we used the TLS feature sets that were studied in previous studies [4,46,8,48,49], which have achieved significantly high classification accuracy for TCP traffic, to compare with the multifractal features obtained by our approach with respect to the classification performance. These TLS feature sets are represented as TLS1, TLS2, TLS3, TLS4 and TLS5, respectively. Each classification experiment is conducted by SVMs over 10-fold CV to obtain the average classification accuracy. Fig. 8 shows comparison of the overall accuracy between the TLS feature sets and the multifractal features (represented as MFF20), which were obtained by the proposed approach with the spacing value of orders q equal to 1.0 from data set D_20_15, for the TCP traffic. As shown in Fig. 8, the overall accuracy of the multifractal features was higher than that of the TLS feature sets. This finding proves that the multifractal features obtained by the proposed approach are more robust and discriminative than the TLS features for traffic classification. Furthermore, these TLS feature sets, especially TLS4 and TLS5, did not achieve the promised high accuracy that was described in the previous literature. This finding is a reasonable result because of the following: (1) it can be observed from Fig. 2 that the transmission pattern of some of the traffic classes in our traffic data are very similar, and thus, their TLS features can hardly be classified from one another, and (2) the transport layer port number, which is contained in all of the TLS feature sets, is very ineffective for our traffic data because WWW and HTTP+Flash use the same



(a) Normalized build time of the SVMs for the TLS features and the multifractal features.

(b) Normalized classification speed of the SVMs for the TLS features and the multifractal features.

Fig. 9. Comparison of the build time and classification speed between the TLS features and multifractal features.

port number of 80, and P2P and IM use dynamic port numbers for communication.

4.6.2. Runtime performance comparison

This section aims to analyze the build time and classification speed for any of the ML algorithms (e. g., SVMs) when using the multifractal features obtained by the proposed approach and the TLS features studied in the previous literature. This is especially important because the model building phase is computationally time consuming and the classification speed affects the efficiency of the traffic classification. Therefore, we use the multifractal features and the TLS feature sets mentioned in the previous section to measure the execution time of building the classification model and the corresponding classification speed using SVMs on a Core i5 2.4 GHz Intel processor machine with 4 Gbytes of RAM. For each of the feature sets, the test was repeated ten times to give the average execution time and classification speed, to achieve greater confidence in the results.

Fig. 9a shows the normalized build time for the SVMs when using the multifractal features in comparison to the TLS features. It can be observed that the time required by TLS4 and TLS5 is quite low. This finding occurs because each of these feature sets has less than 6 features, which can efficiently reduce the build time. It is also notable that the time cost of TLS1 is very high compared to the other feature sets because it has 248 features. Furthermore, it can be found that the build time of the multifractal features is slightly higher than that of TLS2 and TLS3.

Fig. 9b illustrates the classification speed of the SVMs for the traffic flows based on the multifractal features and the TLS features. This finding is especially important when considering real-time classification of potentially thousands of simultaneous network flows. The results show that TLS4 and TLS5 achieved the highest classification speeds, and TLS1 has the lowest classification speed. The multifractal features achieved relatively lower classification speed, which is slightly lower than TLS2 and TLS3.

By comparing Fig. 9a and b, the multifractal features achieved relatively lower runtime performance. However, it should be noted that the TLS features cannot be obtained before the transmission of all of the packets, but the multifractal features can be extracted at the early stage of transmission. For example, the average duration of the P2P flows in our traffic data is approximately 26 s, and the multifractal features in our paper can be extracted at 15 s, and thus, the multifractal features obtained by the proposed approach can achieve earlier traffic classification than the TLS features. Furthermore, parallel computing techniques and multi-core machines can help to make up for the deficiencies in runtime performance of the multifractal features. Therefore, the multifractal features are more effective for real-time traffic classification than the TLS features.

5. Conclusions

In this study, a novel feature extraction and selection approach was proposed to provide the optimal and robust features for traffic classification, where the wavelet leaders based multifractal features are extracted to characterize the traffic flows and the PCABFS method was proposed to remove the irrelevant and redundant features. Based on the observation of the multifractal features and the analysis of PCABFS, the larger Holder exponents and corresponding multifractal spectra were considered to be the optimal features for traffic classification.

To evaluate the classification performance of the proposed approach, the SVMs classifier was employed to evaluate the classification performance of the proposed approach on nine data sets that had different sampling intervals and sampling durations, each of which includes seven different classes of TCP flows and three different classes of UDP flows. The results show that for the TCP flows, the proposed approach achieves the highest classification accuracy for the data set D_20_15, while for the UDP flows, the proposed approach achieves the highest classification accuracy for the data set D_50_20. Therefore, the sampling interval and sampling duration were considered to have a large influence on the classification accuracy of the proposed approach. Furthermore, the spacing value of orders q, which determines the number of multifractal features, also had a slight influence on the classification accuracy, and the overall accuracy of the proposed approach had a slight improvement with an increase in the number of the multifractal features. Moreover, this paper proved the effectiveness of PCABFS method by comparing the accuracy and stability of PCABFS method with those of existing FS methods for different numbers of original multifractal features.

To further demonstrate the performance of the proposed approach, this paper provided two comparisons of classification performance and runtime performance between the multifractal features obtained by the proposed approach and the TLS features studied in the previous literature. The comparison results show that the multifractal features achieved the best classification performance over the TLS features, which proved that the multifractal features are more robust and discriminative than the TLS features. Furthermore, the comparison results of the runtime performance show that the multifractal features obtained relative low runtime performance compared to the TLS features studied in the previous literature. However, due to the ability of classifying traffic at the early stage, the multifractal features are more effective for real-time traffic classification than the TLS features. In conclusion, the multifractal features obtained by the proposed approach not only can achieve higher classification performance than TLS features but also are very suitable for real-time traffic classification. The future work will be devoted to apply the proposed approach into network engineering.

References

[1] T. Auld, A. Moore, S. Gull, Bayesian neural networks for Internet traffic classification,, IEEE Trans. Neural Netw. 18 (1) (2007) 223–239.

[2] F. Hao, M. Kodialam, T. Lakshman, H. Song, Fast dynamic multiple-set membership testing using combinatorial bloom filters, IEEE/ACM Trans. Networking (TON) 99 (2012) 295–304.

[3] S. Sen, O. Spatscheck, D. Wang, Accurate, scalable in network identification of p2p traffic using application signatures, in: Proceedings of the 13th International Conference on World Wide Web, ACM, New York, 2004, pp. 512–521.

[4] A. Moore, K. Papagiannaki, Toward the accurate identification of network applications, in: Passive and Active Network Measurement, 2005, pp. 41–54.

[5] P. Haffner, S. Sen, O. Spatscheck, D. Wang, ACAS: automated construction of application signatures, in: Proceedings of the 2005 ACM SIGCOMM Workshop on Mining Network Data, ACM, 2005, pp. 197–202.

[6] H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, K. Lee, Internet traffic classification demystified: myths, caveats, and the best practices, in: Proc. 4th ACM Int. Conference on Emerging Networking Experiments and Technologies (CoNEXT 2008), Madrid, Spain, 2008, pp. 1–12.

[7] T.T.T. Nguyen, G. Armitage, A survey of techniques for internet traffic classification using machine learning, IEEE Commun. Surv. Tutorials 10 (2008) 56–76.

[8] M. Soysal, E.G. Schmidt, Machine learning algorithms for accurate flow-based network traffic classification: evaluation and comparison, Perform. Eval. 67 (2010) 451–467.

[9] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

[10] H. Liu, H. Motoda, L. Yu, A selective sampling approach to active feature selection, Artif. Intell. 159 (1) (2004) 49–74.

[11] A.W. Moore, D. Zuev, Internet traffic classification using Bayesian analysis techniques, in: Proc. ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, Banff, Alberta, Canada, 2005, pp. 50–60.

[12] N. Williams, S. Zander, G. Armitage, A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification, ACM SIGCOMM Comput. Commun. Rev. 36 (5) (2006) 5–16.

[13] R. Yuan, Z. Li, X. Guan, L. Xu, An SVM-based machine learning method for accurate internet traffic classification, Inf. Syst. Front. 12 (2) (2010) 149–156.

[14] J. Han, M. Kamber, Data Mining, Concepts and Techniques, Morgan, Kaufmann, 2006.

[15] C. Su, J. Hsu, An extended Chi2 algorithm for discretization of real value attributes, IEEE Trans. Knowl. Data Eng. 17 (3) (2005) 437–441.

[16] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second edition, John Wiley & Sons, New York, 2001.

[17] M. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in: Proceedings of the 7th International Conference on Machine Learning, ACM, 2000, pp. 359–366.

[18] L. Yu, H. Liu, Feature selection for high-dimensional data: a fast correlation-based filter solution, in: Proceedings of the 20th International Conference on Machine Learning, ICML, 2003, pp. 856–863.

[19] A.C. Gilbert, W. Willinger, A. Feldmann, Scaling analysis of conservative cascades, with applications to network traffic, IEEE Trans. Inf. Theory 45 (3) (1999) 971–991.

[20] S. Molnár, I. Maricza, Source Characterization In Broadband Networks, Interim Report, COST 257, Vilamoura, Portugal, January 1999.

[21] R.H. Riedi, M.S. Crouse, V.J. Ribeiro, R.G. Baraniuk, A multifractal wavelet model with application to network traffic, IEEE Trans. Inf. Theory 45 (3) (1999) 992–1018.

[22] R.H. Riedi, W. Willinger, oward an Improved Understanding of Network Traffic Dynamics, Self-Similar Network Traffic and Performance Evaluation, in: K. Park, W. Willinger (Eds.), Wiley, June 1999.

[23] J. Beran, in: Statistics For Long-Memory Processes, Chapman & Hall, One Penn Plaza, New York, NY, 1995, p. 10119.

[24] P. Abry, D. Veitch, Wavelet analysis of long range dependent traffic, IEEE Trans. Inf. Theory 44 (1) (January 1998) 2–15.

[25] M.S. Taqqu, V. Teverovsky, W. Willinger, Is network traffic self-similar or multifractal? Fractals 5 (1997) 63–73.

[26] A. Feldmann, A.C. Gilbert, W. Willinger, Data networks as cascades: investigating the multifractal nature of Internet WAN traffic, ACM Comput. Commun. Rev. 28 (1998) 42–55.

[27] A. Feldmann, A.C. Gilbert, W. Willinger, T.G. Kurtz, The changing nature of network traffic: scaling phenomena, ACM Comput. Commun. Rev. 28 (1998) 5–29.

[28] E. Bacry, J. Muzy, A. Arnéodo, Singularity spectrum of fractal signals from wavelet analysis: exact results, J. Stat. Phys. 70 (3) (1993) 635–674.

[29] R. Lopes, N. Betrouni, Fractal and multifractal analysis: a review, Med. Image Anal. 13 (4) (2009) 634–649.

[30] D. Lim, Principal component analysis using singular value decomposition of microarray data, Int. J. Math. Comput. Phys. Quantum Eng. 7 (9) (2013) 859–861.

[31] J. Yang, PCA based sequential feature space learning for gene selection, in: Proceedings of the IEEE International Conference on Machine Learning and Cybernetics, ICMLC, 2010, pp. 3079–3084.

[32] Y. Luo, A PCA based unsupervised feature selection algorithm, in: Proceedings of the IEEE International Conference on Genetic and Evolutionary Computing, WGEC, 2008, pp. 299–302.

[33] F. Song, Feature selection using principal component analysis, in: Proceedings of the IEEE System Science, Engineering Design and Manufacturing Informatization, ICSEM, 2010, pp. 27–30.

[34] T. Karagiannis, K. Papagiannaki, M. Faloutsos BLINC, Multilevel traffic classification in the dark, SIGCOMM Comput. Commun. Rev. 35 (4) (2005) 229 240.

[35] T. Karagiannis, A. Broido, M. Faloutsos, K.C. Claffy, Transport layer identification of P2P traffic, 4th ACM SIGCOMM Conference on Internet Measurement, 2004.

[36] K. Wang, S.J. Stolfo, Anomalous Payload-Based Network Intrusion Detection, Springer, Berlin, 2004.

[37] Snort [Online], Available: http://www.snort.org/.

[38] S. Yi, B.-K. Kim, J. Oh, J. Jang, G. Kesidis, C.R. Das, Memory-efficient content filtering hardware for high-speed intrusion detection systems, in: ACM Symposium on Applied Computing, 2007, pp. 264–269.

[39] The Internet Assigned Numbers Authority, IANA, [Online], Available: http://www.iana.org/.

[40] D. Moore, K. Keys, R. Koga, E. Lagache, K.C. Claffy, The CoralReef software suite as a tool for system and network administrators, in: 15th USENIX Conference on System Administration, LISA01, pp. 133–144.

[41] M. Crotti, F. Gringoli, P. Pelosato, L. Salgarelli, A statistical approach to IP-level classification of network traffic, in: IEEE International Conference on Communications, 2006, pp. 170–1765.

[42] M. Roughan, S. Sen, O. Spatscheck, N. Duffield, Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification, in: 4th ACM SIGCOMM Conference on Internet Measurement, 2004, pp. 135–148.

[43] A. McGregor, M. Hall, P. Lorier, J. Bruuskill, Flow clustering using machine learning techniques, in: Proceedings of the 5th Passive and Active Measurement Workshop (PAM), France, Antibes Juanles-Pins, 2004, pp. 205–214.

[44] S. Zander, T. Nguyen, G. Armitage, Automated traffic classification and application identification using machine learning, in: Proceedings of IEEE Conference on Local Computer Networks, Sydney, 2005, pp. 250–257.

[45] E.G. Schmidt, M. Soysal, An intrusion detection based approach for the scalable detection of P2P traffic in the national academic network backbone, in: International Symposium on Computer Networks, 2006, pp. 128–133.

[46] W. Li, M. Canini, A.W. Moore, Efficient application identification and the temporal and spatial stability of classification schema, Comput. Netw. 53 (6) (2009) 790–809.

[47] L. Dai, X.C. Yun, J. Xiao, Optimizing traffic classification using hybrid feature selection, The Ninth International Conference on Web-Age Information Management, Hunan, China, July 2008.

[48] H.L. Zhang, G. Lu, M.T. Qassrawi, Y. Zhang, X.Z. Yu, Feature selection for optimizing traffic classification, Comput. Commun. 35 (12) (2012) 1457–1471.

[49] A. Fahad, Z. Tari, I. Khalil, A. Almalawi, An optimal and stable feature selection approach for traffic classification based on multi-criterion fusion, Future Gener. Comput. Syst. 36 (2014) 156–169.

[50] M. Soysal, E.G. Schmidt, An accurate evaluation of machine learning algorithms for flow-based P2P traffic detection, 22nd International Symposium on Computer and Information Sciences, 2007 ISCIS07.

[51] N. Williams, S. Zander, G. Armitage, Evaluating machine learning algorithms for automated network application identification, Center for Advanced Internet Architectures, CAIA, 2006 Technical Report 060410B.

[52] A. Arneodo, G. Grasseau, M. Holschneider, Wavelet transform of multifractals, Phys. Rev. Lett. 61 (20) (1988) 2281–2284.

[53] J.F. Muzy, E. Bacry, A. Arnéodo, Multifractal formalism for fractal signals: the structure-function approach versus the wavelet-transform modulusmaxima method, Phys. Rev. E 47 (2) (1993) 875–884.

[54] H. Shi, G. Liang, H. Wang, A novel traffic identification approach based on multifractal analysis and combined neural network, Ann. Telecommun. 69 (2014) 155–169.

[55] Z. Peng, F. Chu, P.W. Tse, Singularity analysis of the vibration signals by means of wavelet modulus maximal method, Mech. Syst. Signal Process. 21 (2) (2007) 780–794.

[56] S. Mallat, A Wavelet Tour of Signal Processing, San Diego, CA: Academic, 1998.

[57] H. Wendt, P. Abry, S. Jaffard, Bootstrap for empirical multifractal analysis, IEEE Signal Process. Mag. 24 (4) (2007) 38–48.

[58] A.B. Chhabra, C. Meneveau, R.V. Jensen, K. Sreenivasan, Direct determination of the f(α) singularity spectrum and its application to fully developed turbulence, Phys. Rev. A 40 (9) (1989) 52–84.

[59] S. Jaffard, Multifractal formalism for functions. Part I: results valid for all functions, SIAM J. Math. Anal. 28 (4) (1997) 944–970.

[60] G.H. Golub, C. Reinsch, Singular value decomposition and least squares solutions, Numer. Math. 14 (1970) 403–420.

[61] H.F. Kaiser, A note on Guttman's lower bound for the number of common factors1, Br. J. Stat. Psychol. 14 (1961) 1–2.

[62] H.F. Kaiser, The varimax criterion for analytic rotation in factor analysis, Psychometrika 23 (1958) 187–200.

[63] H. Wendt, P. Abry, S. Jaffard, et al., Wavelet leader multifractal analysis for texture classification, in: Proceedings of the IEEE International Conference on Image Processing, ICIP, 2009, pp. 3829–3832.

[64] H. Wendt, SG. Roux, S. Jaffard, P. Abry, Wavelet leaders and bootstrap for multifractal analysis of images, Signal Process. 89 (6) (2009) 1100–1114.

[65] P. Abry, H. Wendt, S. Jaffard, et al., Methodology for multifractal analysis of heart rate variability: from LF/HF ratio to wavelet leaders, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10, 2010, pp. 106–109.

[66] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceeding of the 1992 Fifth Annual ACM Workshop, 1992, pp. 144–152.

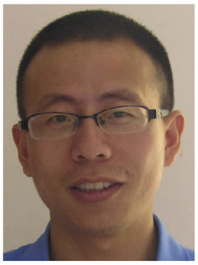
[67] C.W Hsu, C.J. Lin, A comparison of methods for multiclass support vector machines, IEEE Trans. Neural Netw. 13 (2) (2002) 415–425.

[68] T Hastie, R Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second edition, Springer, New York, USA, 2009.

[69] S. Mitra, Y Hayashi, Bioinformatics with soft computing, IEEE Trans. Syst. Man Cybern. Part C 36 (5) (2006) 616–635.

[70] A. Kalousis, J. Prados, M. Hilario, Stability of feature selection algorithms: a study on high-dimensional spaces, Knowl. Inf. Syst. 12 (1) (2007) 95–116.

[71] P. Drotár, J. Gazda, Z. Smékal, An experimental comparison of feature selection methods on two-class biomedical datasets, Comput. Biol. Med. 66 (2015) 1–10.



Hongtao Shi is currently a Ph. D. candidate in College of Information Science and Engineering at Ocean University of China. He received his B.S. in Computer Sciences from Chang'an University, and his M.S. in Computer Application from Ocean University of China. His research interests include traffic measurement, characterization and analysis, performance evaluation, traffic engineering, security and network design.

Hongping Li is currently a professor in College of Information Science and Engineering at Ocean University of China. He received his Ph. D. in Computer Sciences from University of Oklahoma. His research interests include machine learning, feature selection, traffic management.

Dan Zhang is currently a graduate student in College of Information Science and Engineering at Ocean University of China. Her research interests include data mining, machine learning and big data.

Chaqiu Cheng is currently a graduate student in College of Information Science and Engineering at Ocean University of China. His research interests include data mining, statistical analysis and machine learning.

Wei Wu is currently a Ph. D. candidate in College of Information Science and Engineering at Ocean University of China. He received his B.S. in Computer Sciences from Shandong Agricultural University, and his M.S. in Computer Application from Ocean University of China. His research interests include computer vision, image analysis and machine learning.