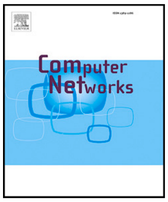
Computer Networks 185 (2021) 107645

Contents lists available at ScienceDirect

Computer Networks

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



A multi-channel anomaly detection method with feature selection and multi-scale analysis

Lisheng Huang [a,1], Jinye Ran [b,1], Wenyong Wang [a,1], Tan Yang [c], Yu Xiang [a,*]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 61173, China
[b] Operations Management Department, China Merchants Bank, Shenzhen 518040, China
[c] School of Software Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

ARTICLE INFO

Keywords:
Network traffic
Multi-channel GLRT
Anomaly detection
EEMD
Feature selection

ABSTRACT

This paper proposes a novel multi-channel network traffic anomaly detection method combined with the idea of multi-scale decomposition, feature selection and fusion. Our method includes a feature selection and fusion module, which extracts the most important features from redundancy traffic features and fuses the selected features, a multi-scale decomposition module, and a multi-channel Generalized Likelihood Ratio Test (GLRT) detection module, for anomaly detection and decision-making. The advantage of our method is that, first, it only considers the selected features and reduces the amount of computation. Second, compared with traditional anomaly detection methods that usually work on each scale independently thus mainly focused on temporally correlated traffic, our method fully explores the internal frequency–time correlations within multiple scales. It can be verified with experiments that this method performs better than other traditional methods, thus gives a new sight on the anomaly detection with different types of traffic data.

1. Introduction

Anomaly detection in computer network traffic faces challenges with the cyber-attacks increasing. With the digitalization, it is becoming more and more important to ensure the security of Internet users' data, at the same time, more and more computer viruses are spreading in the network, and various forms of cyber-attacks emerge in an endless stream [1].

Network traffic anomaly detection is an important research area in network security that detects anomalous or abnormal network traffic data from a given dataset. It is an interesting research area that many works have been proposed for solving this problem. Despite the large number of techniques available, using signal processing technique for anomaly detection has hardly been explored. In [2], the main challenges of network traffic anomaly detection is summarized as following.

- A lack of universally applicable anomaly detection technique;
- Network traffic data usually contains noise that makes the anomaly detection task more difficult.
- A lack of publicly available labeled dataset to be used for network anomaly detection.
- As normal behaviors are continually evolving and may not be normal forever, current intrusion detection techniques may not be useful in the future.

Due to the above-mentioned challenges, the network traffic anomaly detection has become a challenging task, and therefore the research community has increased interest about this research area.

The main contribution of this paper is to propose a novel network traffic anomaly detection method with multi-scale decomposition and multi-channel anomaly detection. Our method firstly introduces newly-developed multi-channel anomaly detection theory into traffic anomaly detection field. And the adoption of our method not only explores the temporal relationship of network traffic, but also comprehensively explores the internal frequency–time correlations among different scales during the detection process. Meanwhile, our method also combines the idea of feature selection, which significantly reduces the amount of computation by extracting the most important features from redundancy traffic features, and makes our method achieves better performance than other methods.

The rest of the paper is organized as follows. The related work is described in Section 2. The system model is mainly introduced in Section 3. Section 4 shows the experimental results and analysis. Then, the main conclusions are summarized in Section 5.

[*] Corresponding author.
E-mail address: jcxiang@uestc.edu.cn (Y. Xiang).

[1] These authors contributed to the work equally and should be regarded as co-first authors.

https://doi.org/10.1016/j.comnet.2020.107645
Received 18 October 2019; Received in revised form 1 September 2020; Accepted 19 October 2020
Available online 22 October 2020

2. Related works

Researchers have approached anomaly detection using various techniques. The statistics community has been studying the problem of detection of anomalies or outliers from as early as the 19th century [3]. Statistically speaking, an anomaly, also called an outlier, usually can be defined as an observation which is suspected of being partially or wholly irrelevant because it is not generated by the stochastic model assumed [4]. The statistical-based anomaly detection technology is one of the most well researched technology compared with other technologies. In many network anomaly detection systems, it was used as one of the key technologies. Basically, statistical-based methods fit a statistical model for normal behavior under a given distribution assumption, and then apply a statistical inference test to determine if an unseen instance belongs to this model. In [5], Eskin et al. proposed a general mixture model based on a majority distribution and an anomaly distribution. However, in order to learn both models, data of two distributions are required. Laxhammar et al. [6] proposed a statistical method using a probabilist threshold ϵ to discriminate normal or anomalous data. Like many other statistical methods [7–9], these techniques typically suffer from one or more limitations, such as assuming that they are too strict to address complex problems, and need a huge training sample if we want a low false positive rate.

Recently, artificial intelligence related methods (such as supervised/unsupervised machine learning algorithms, deep neural networks) [2,10–14] have started to play an important role in anomaly detection. Basically these methods treat anomaly detection as a classification/clustering problem. These methods usually face the challenge of the imbalanced proportion of normal and anomalous samples. Due to the inherent complex characteristics of imbalanced data sets, learning from such data requires new understandings, principles, algorithms, and tools to transform vast amounts of raw data efficiently into information and knowledge representation [15]. Some state-of-the-art anomaly detection methods was based on the organic integration of multiple artificial intelligence technologies, which can be well adapted to the complex and truly network attack environment. In [16], Tama et al. proposed a two-stage classifier ensemble based on rotation forest and bagging for anomaly detection. After that, Tama et al. [17] proposed a stacked ensemble for anomaly detection method that compose of random forest, gradient boosting machine and XGBoost. This method solved the problem of imbalanced data sets by ensemble learning. Ying et al. [18] applied heterogeneous ensemble learning in anomaly detection, they combined the unsupervised Autoencoder with the supervised Long Short-Term Memory. The above ensemble methods used a dataset to train different learners to detect anomaly. These give us a new thought to use multiple decomposed subdatasets to reduce the influence of imbalance of initial dataset. We treat traffic data as temporally related signals that we could decompose the initial dataset into multiple subgroups, then we analysis anomaly using these subgroups.

Although signal processing is an interesting research area, using such a technique for anomaly detection has hardly been explored [2]. Network traffic data can be viewed intuitively as a time series. Thus, the anomaly detection can be regarded as signal detection from a signal processing point of view. Thottan et al. believed that signal processing technique has great potential to enhance the field [19]. Daniela et al. by combining network topology information and transform-domain analysis, revisited principal component analysis (PCA) based anomaly detection approaches [20]. Ren et al. proposed a algorithm that dividing time series into several subsequences, based on similarity results among subsequences, the anomaly was detected [21]. Kaloorazi et al. proposed randomized subspace methods to detect anomalies in network [22]. Transform-domain based anomaly detection was proposed and explored in [23]. Although signal processing methods have made great progress in anomaly detection fields, it still needs to be improved.

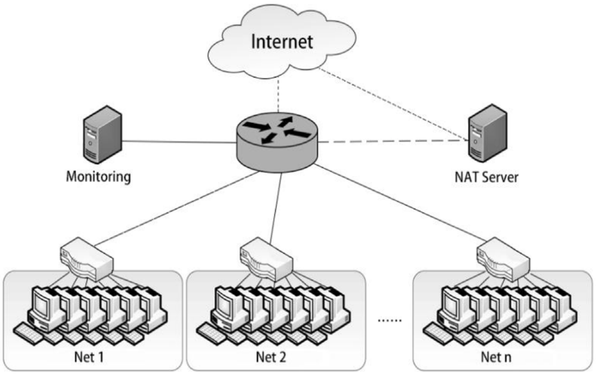


Fig. 1. Network architecture.

Compared with traditional signal processing based anomaly detection methods, multi-scale analysis has been proposed as an effective way in anomaly detection, and researchers have shown that, multi-channel detection that could capture the spatial–temporal correlation among multiple data sets performs better with considered [24,25]. Jiang D. et al. used PCA with continuous wavelet transforms to extract the nature of the anomalous network traffic [26]. R. Fontugne et al. presented a new traffic anomaly detector called sketch and multi-scale (SMS), which combined random projections (sketches) with multi-scale analysis based on wavelet [27]. Captured network traffic represents the behavior characteristics of the traffic from a given network topology. As shown in Fig. 1, in such a given network topology, traffic data is acquired on predetermined nodes (always on boundary/backbone routers). Paschalidis I. et al. considered the difference in spatial location between different nodes, and analyzed the spatio-temporal correlation of traffic data according to the relationship between different time and node data, as a result, the traffic data acquired is spatio-temporal correlated, with model-free and MMP model-based approaches, which use some spatial information to expand existed temporal-correlated algorithms [28]. Cross entropy method, together with likelihood ratio test, were adopted in [29] to detect anomalies in temporally correlated traffic. Researchers have shown that, compared with single-channel detection, multi-channel detection that could capture the spatial–temporal correlation among multiple data sets performs better with considered [24,25]. But multi-channel detectors, while own better performance, they have not been applied to network anomaly detection to the best of our knowledge. Together with multi-scale data decomposition such as wavelet and EMD, traffic data can be decomposed into multiple scales and anomalies could be detected on one or more timescales components [26,27,30]. Although many methods for anomaly detection which combined with multi-scale analysis of traffic characteristics were already proposed, they treated each scale independently, combined individual detection decisions into a single ‘global decision’, regardless of the correlations among the observations.

In the previous multi-channel anomaly detection research, researchers focused on wavelet transform. In recent years, N.E Huang et al. proposed a new multi-scale decomposition method named ensemble empirical mode decomposition (EEMD) [31]. Unlike wavelet transform that based on a priori wavelet basis function, EEMD can adaptively decompose nonlinear and non-stationary data into multiple components. Due to this advantage, EEMD has achieved good effects in many field including signal processing and detection [2,32,33], but it has not been widely explored in network anomaly detection field.

3. System model

The framework of our proposed multi-channel anomaly detection method is shown in Fig. 2. The multi-channel anomaly detection model can be divided into three modules: feature selection and fusion module, multi-scale decomposition module and multi-channel anomaly detection module, each module will be introduced in detail as follow.

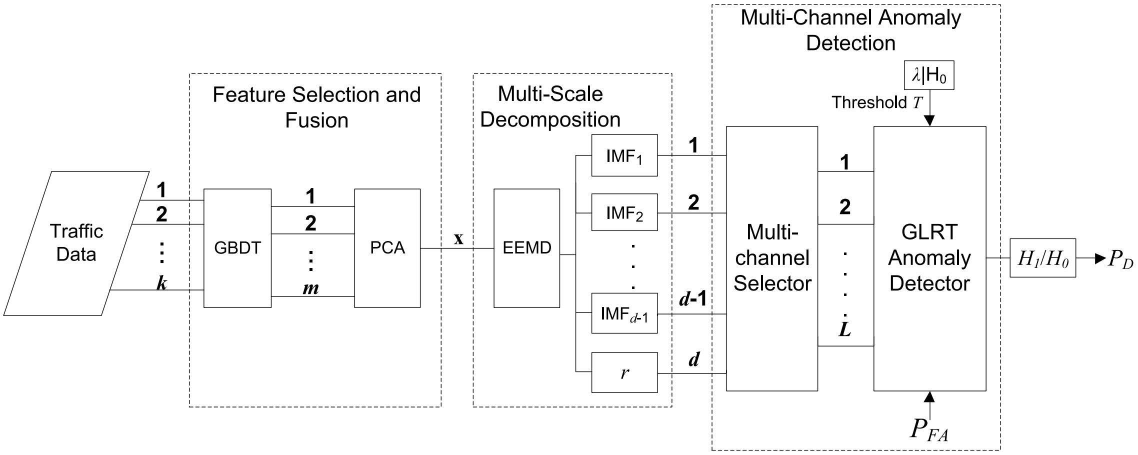


Fig. 2. The framework of multi-channel anomaly detector.

3.1. Feature selection and fusion

After processing the original traffic data, network traffic features with standard definitions can be acquired, such as downlink or uplink throughput, number of downlink or uplink packet, server or client jitter, etc., see Table 2. We define the processed traffic data as the traffic feature vector, which is of high dimensional. However, In the scenario of traffic anomaly detection, not all features have a high correlation with anomalies, that is, features in the traffic data are usually highly redundant. High dimensional data and redundant features will not only affect the efficiency of the processing, but also affect the accuracy of the calculation, so it makes identifying informative features a necessary step for effective processing and analysis in traffic data anomaly detection.

In this paper, we select features using gradient boosting decision tree (GBDT) [34], which was developed to identify the features of users' comments about items, and this method tends to select features that have the highest correlation with network anomaly. After selecting the most important features, we use PCA to fuse the information of multiple features, and finally acquire the data with smaller dimension.

3.1.1. Feature selection

Feature selection is to select a subset from the feature space. Feature selection methods can be categorized into three types depending on their selection mechanism: filters, wrappers, and embedded. Filter methods use general characteristics such as correlation to remove irrelevant features. Wrapper methods use classifiers or specific target functions to evaluate the performance and to search for the best combination of features. Embedded methods are similar to wrappers that they also use a classifiers but they perform feature selection as a part of the classification process. The embedded methods can have a better performance in the complex feature selection tasks, and the get a more accurate result, which is exactly selected in our method.

As a embedded feature selection method, GBDT perform feature selection in the process of binary classification, the weak classifiers-decision trees is capable of calculating the feature importance based on information gain. The basic idea of GBDT is to combine a series of weak learners into a stronger one, and the weak learner used here is decision tree. Different from the traditional boosting methods that weight positive and negative samples, GBDT makes global convergence of algorithm by following the direction of the negative gradient [35].

GBDT is often used as a classifier, however, in this paper, we use GBDT for feature selection. Basically, GBDT is an iterative classification model, each iteration calculates the information gain of different features, and each time selects the feature with the largest information gain as the leaf node, which we will describe later. Therefore, we consider that the binary classification process of GBDT also is feature selection process.

The main idea behind GBDT uses binary classification, in which a scalar score function is formed to distinguish the two classes. Let $D = \{\mathbf{t}_i, y_i\}_{i=1}^{n}$ denotes the dataset, where $\mathbf{t}_i \in R^k, \mathbf{t}_i = (t_{i1}, t_{i2}, \dots, t_{ik})$, represents a traffic feature vector with k features, y_i is the predicted label with $y_i \in \{0, 1\}$, indicates whether the traffic is anomaly. The steps of GBDT are presented as follows:

1. Given the initial constant value to the model β

$$F_0(\mathbf{t}) = \arg\min_{\beta} \sum_{i=1}^{N} Loss\left(y_i, \beta\right) \tag{1}$$

2. For the ith iteration, where $k = 1 : I$, and I is the times of iteration, the gradient of residuals is acquired.

$$y_i^* = -\left(\frac{\partial Loss\left(y_i, F\left(\mathbf{t}_i\right)\right)}{\partial F\left(\mathbf{t}_i\right)}\right)_{F(\mathbf{t})-F_{\mathrm{k}-1}(\mathbf{t})}, \tag{2}$$

where $i = \{1, 2, \dots, N\}$.

3. The basic classifiers are used to fit sample data and get the model. According to the least square approach, parameter a_k of the model is obtained and the model $h(x_i; a_k)$ is fitted.

$$a_k = \arg\min_{\beta} \sum_{i=1}^{N} \left[y_i^* - \beta h\left(\mathbf{t}_i; a\right)\right]^2 \tag{3}$$

4. Minimize the loss function $Loss(y_i, F(\mathbf{t}_i))$. And calculate the weight of new model.

$$\beta_k = \arg\min_{\beta} \sum_{i=1}^{N} Loss\left(y_i, F_{k-1}(\mathbf{t}) + \beta h\left(\mathbf{t}_i; a\right)\right) \tag{4}$$

5. Update the model.

$$F_k(\mathbf{t}) = F_{k-1}(\mathbf{t}) + \beta_k h\left(\mathbf{t}_i; a\right) \tag{5}$$

In summary, after I iterations, the function of GBDT can be formed in

$$F(\mathbf{t}) = \sum_{k=1}^{I} F_k(\mathbf{t}) \tag{6}$$

Note that, the process of generating basic classifier in each iteration, namely a decision tree, is actually a process of feature selection. As shown in Fig. 3, each node in a decision tree selected a feature point with the best classification effect, and the feature selection scheme will be described in detail as follow.

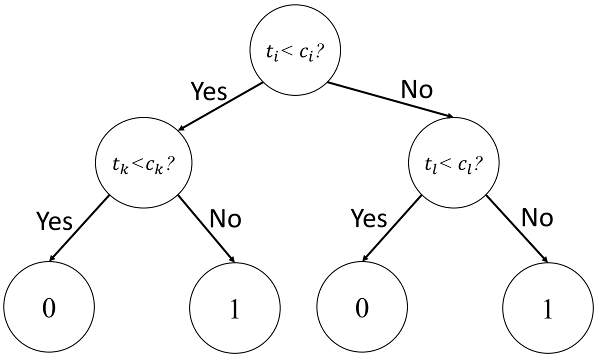


Fig. 3. Decision tree.

Let $H(D)$ denotes the entropy of data set D, which is defined as

$$H(D) = -\sum_{i=0}^{1} p_i \log p_i \tag{7}$$

where p_0 is probability density of normal traffic, and p_1 is probability density of anomalous traffic. Let $H(D|A)$ denotes conditional entropy of D under a given feature A

$$H(D|A) = \sum_{i=1}^{n} p_i' H(Y|A = A_i) \tag{8}$$

where p_i' is the probability density when $A = A_i$. The information gain of feature A $g(D, A)$ is defined as

$$g(D, A) = H(D) - H(D|A) \tag{9}$$

It can be known from Eq. (9) that the larger the information gain, the better the effect of selecting the feature to classify the dataset. For each iteration, the information gains of all features are calculated, and the features with the largest information gain are selected as the split nodes to generate a new decision tree. After the GBDT model is finished training, in theory, the greater the correlation between the feature and anomaly, the more times the feature is selected as the split nodes. So the number of times a feature is selected as the split nodes can be used as a measure to select the highest features. After feature selection, the new traffic feature vector is

$$\mathbf{t'}_i = (t_{is_1}, t_{is_2}, \dots, t_{is_m}) \tag{10}$$

where $s_1 < s_2 < \cdots < s_m$ and $m \leq k$.

3.1.2. Fusion

In this article, we aim to re-fuse features to get lower-dimensional data. It is similar to a dimensionality reduction process. There are several approaches have been proposed to achieve this goal.

As an excellent dimensionality reduction method, Linear Discriminant Analysis (LDA) is widely used in various applications. But for the task of network traffic features fusion, this method has some limitations as follow, (1) LDA is a supervised dimensionality reduction method, but feature fusion is an unsupervised dimensionality reduction task; (2) LDA can only reduce the original data to $k - 1$ dimension; (3) LDA is not suitable for dimensionality reduction of non-Gaussian distribution samples.

In recent years, autoencoder has also been widely used as a dimensionality reduction method, but because this method is based on neural network technology, requires a large amount of training data, and is not a general dimensionality reduction technology, thus it cannot be applied to the application scenarios of this article.

In this paper, we use PCA to fuse multidimensional traffic feature vector into one-dimensional time series. PCA is a multivariate technique that analyzes a data table in which observations are described by several inter-correlated quantitative dependent variables. Its goal is to extract the important information from the table, to represent it as a set of new orthogonal variables called principal components [36,37].

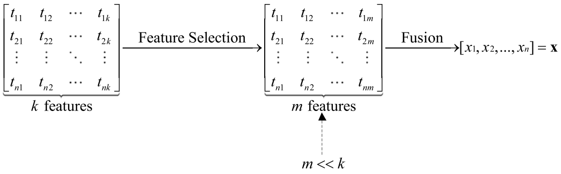

t_{11} t_{12} $\cdots$ t_{1k}
t_{21} t_{22} $\cdots$ t_{2k}
t_{n1} t_{n2} $\cdots$ t_{nk}
k features
Feature Selection
t_{11} t_{12} $\cdots$ t_{1m}
t_{21} t_{22} $\cdots$ t_{2m}
t_{n1} t_{n2} $\cdots$ t_{nm}
m features
$m << k$
Fusion
$[x_1, x_2, \dots, x_n] = \mathbf{x}$

Fig. 4. Feature selection and fusion.

Taking $\mathbf{t'}_i$ in Eq. (10) as an example, the main steps of PCA are as follows:

1. Preprocessing data, make sure the columns of $\mathbf{t'}_i$ is centered so that the mean of each column is equal to 0;
2. Calculating the covariance matrix $\mathbf{C} = \mathbf{t'}_i^T \mathbf{t'}_i$;
3. Calculating the eigenvalues $\{e_i\}$ of the $\mathbf{C}$ and the orthogonal eigenvectors $\{v_i\}$;
4. Taking the eigenvectors corresponding to the first $n (n \leq m)$ eigenvalues to obtain the matrix $\mathbf{P} = [e_1, e_2, \dots, e_n]^T$;
5. Letting $\mathbf{x} = \mathbf{P}\mathbf{t}_i'$, where $\mathbf{t}$ is the compressed data.

Note that in order to fuse the multidimensional traffic feature vector into a one-dimensional time series, the value of n in Step 3 above is 1.

3.1.3. Feature selection and fusion

This section will introduce the process of feature selection and fusion of the method in this paper. As shown in Fig. 4, this method is based primarily on GBDT and PCA, the main steps are as follows:

1. Using the labeled traffic data $D = \{\mathbf{t}_i, y_i\}_{i=1}^n, \mathbf{t}_i \in R^k, \mathbf{t}_i = (\mathbf{t}_{i1}, \mathbf{t}_{i2}, \dots, \mathbf{t}_{ik})$, as input, follow the steps in 2.1.1 to generate a GBDT model;
2. According to the GBDT model, select m features with high importance;
3. For all traffic data, only the features selected in 3 are retained, and $D' = \{\mathbf{t'}_i\}_{i=1}^n, \mathbf{t'}_i \in R^m$, is obtained;
4. Using D' as input, follow the steps in 2.1.2, the selected and fused data $\mathbf{x}$ is obtained.

Note that, from the above we know that both the feature fusion method and feature selection proposed in this paper can be understood as data dimensionality reduction methods. But there are essential differences between them.

Feature selection only filters the original features and selects several features from them without any changes to these features. There will not be any new features in this process, but some original "redundant/wasted" features will be discarded.

Compared with feature selection, PCA is to construct a mapping from high-dimensional to low-dimensional based on the idea of maximizing variance. In this process, PCA will remap the original features in the new space and generate new features. The new feature is a representation of the original features in the low-dimensional space. It retains as much information as possible in the original data. Therefore, unlike feature selection, we call this process feature fusion. Additionally, the process of feature fusion can make the data format meet the requirements of subsequent anomaly detectors and improve the efficiency of subsequent operations. In general, the difference can be summarized as follow:

- Feature fusion can generate new features based on the original features. The new feature is a mapping of the original feature in the new coordinate system that retains as much of the information as possible.

- Feature selection does not make any changes to features, nor does it generate new features, only discarding redundant features.

3.2. Multi-scale decomposition

In order to get the characteristics of network traffic on different time scales, we use EEMD method to decompose network traffic into several components. EEMD is a noise assisted data analysis method, can be used to overcome mode mixing problem caused by signal intermittency in traditional empirical mode decomposition (EMD) [38]. Like EMD, EEMD can also be seemed as a shift progress, which can decompose nonlinear non-stationary signals into d components and a residue r. EEMD consists of sifting an ensemble of white noise-added signal (data) and treats the mean as the final true result. Finite, not infinitesimal, amplitude white noise is necessary to force the ensemble to exhaust all possible solutions in the sifting process, thus making the different scale signals to collate in the proper IMF [31]. It can be defined as follow,

$$\mathbf{x} = \sum_{i=1}^{d} imf_i + r \tag{11}$$

where $\mathbf{x}$ denotes time series data, imf_i denotes the ith intrinsic model function (IMF). The processes of EEMD can be summarized as follows:

1. Adding white noise sequence to the original data to get the sequence, $\mathbf{x}'$;
2. Decomposing sequence $\mathbf{x}'$ into IMFs and a residual;
3. Repeating (1) and (2) with the different white noise added;
4. Calculating the mean value with IMFs and the residual which are decomposed for N times, then getting the final decomposition results.

3.3. Multi-channel anomaly detection

The traffic data after multi-scale decomposition can be formed as follows:

$$\mathbf{X} = \underbrace{\begin{bmatrix} imf_1(0) & imf_1(1) & \cdots & imf_1(N-1) \\ imf_2(0) & imf_2(1) & \cdots & imf_2(N-1) \\ \vdots & \vdots & \ddots & \vdots \\ r(0) & r(1) & \cdots & r(N-1) \end{bmatrix}}_{N \ samples \ times}$$

$$= \left.\begin{bmatrix} \mathbf{imf_1}^{\mathbf{T}} \\ \mathbf{imf_2}^{\mathbf{T}} \\ \vdots \\ \mathbf{imf_{L-1}}^{\mathbf{T}} \\ \mathbf{r}^{\mathbf{T}} \end{bmatrix}\right\} L \ scales$$

The ith row of $\mathbf{X}$ contains information at a specific time scale that is captured of the ith imf or the residual r, and the kth column contains information that is captured within all the time scales at the same time. To make a tradeoff between efficiency and performance, we need to decide the number of detection channel L, which regards components with higher frequencies as independent channels, while other lower-frequency components could be accumulated. Specifically, the number of detection channels is determined by the number of all the decomposed components d, and we choose $L = \lfloor (d-1)/2 \rfloor$. Data segment length N is a pre-determined value in the following GLRT detector, always set from 30 to 50 according to [24,25].

After column vectorization operation, the vector $\mathbf{z} = vec(\mathbf{X}^T)$, and the sample covariance matrix $\mathbf{R}$ could be get as

$$\mathbf{R} = E[\mathbf{z}\mathbf{z}^H] = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{11}^H & \cdots & \mathbf{R}_{L1}^H \\ \mathbf{R}_{21} & \mathbf{R}_{22} & \cdots & \mathbf{R}_{L2}^H \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{L1} & \mathbf{R}_{L2} & \cdots & \mathbf{R}_{LL} \end{bmatrix} \tag{12}$$

Table 1
Definitions of four detection results.

		Detection result	
		T	F
Actual	T	True Positive (TP)	False Negative (FN)
	F	False Positive (FP)	True Negative (TN)

with $\mathbf{R}_{ik} = \mathbf{R}_{ki}^H = E[\mathbf{imf}_i \mathbf{imf}_k^H], 1 \le i,k \le L$. Matrix $\mathbf{R}$ captures all the space–time second-order information for both individual samples and every pair of sample, which is the sample cross-covariance matrix between the ith and kth time series. In multi-channel anomaly detection, we consider the following detection function $\lambda(\mathbf{X}_{1:L}^{(1:t)})$, which maps the observation space $\mathbf{X}_{1:L}^{(1:t)}$ into space $\mathcal{H} = \{\mathcal{H}_0, \mathcal{H}_1\}$:

$$\lambda(\mathbf{X}_{1:L}^{(1:t)}) : \mathbf{X} \to \mathcal{H} \tag{13}$$

As shown in Eq. (13), our problem turns out to be a binary hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, which mean the appearance of normal traffic and anomalous traffic respectively, it can be defined as follows:

$$\begin{cases} \mathcal{H}_0 : \mathbf{x} = \mathbf{x}_N \\ \mathcal{H}_1 : \mathbf{x} = \mathbf{x}_N + \mathbf{x}_A \end{cases} \tag{14}$$

where $\mathbf{x}$ represents the network traffic data, $\mathbf{x}_N$ represents the normal traffic, and $\mathbf{x}_A$ represents the abnormal traffic. To solve this binary hypothesis problem and make a decision, some statistical tests must be applied with a prediction error.

As described by Section 2, although many methods for anomaly detection which combined with multi-scale analysis of traffic characteristics were already proposed, they treated each scale independently, combined individual detection decisions into a single 'global decision', regardless of the correlations among the observations.

3.3.1. GLRT

In [24,25], researchers has proposed a multi-channel GLRT detector, which has been extensively studied in the field of signal processing, but has not yet been applied to network traffic anomaly detection. Multi-channel GLRT detector is able to comprehensively analyze the information on multi-scales and fully considers the internal frequency–time correlations within multiple scales of traffic data.

This paper aims to explore the application of multi-channel GLRT detector in network traffic anomaly detection. In time domain multi-channel GLRT can be described as

$$\begin{aligned} \lambda &= \left(\frac{\max_{R \in \Re_0} p(\mathbf{z}_0, \ldots, \mathbf{z}_{M-1}; \mathbf{R})}{\max_{\mathbf{R} \in \Re_1} p(\mathbf{z}_0, \ldots, \mathbf{z}_{M-1}; \mathbf{R})} \right)^M \\ &= det(\hat{\mathbf{R}}_0^{-1} \hat{\mathbf{R}}_1)^M = \left(\frac{\hat{\mathbf{R}}}{\prod_{i=1}^{L} \det\left(\hat{\mathbf{R}}_{ii}\right)} \right)^M \end{aligned} \tag{15}$$

where $\mathbf{R}_0 \in \Re_0, \mathbf{R}_1 \in \Re_1$ represent the covariance matrix corresponding to $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively, we set $\mathbf{R}_0$ and $\mathbf{R}_1$ are the sets of the block-diagonal covariance matrices in $\mathbf{R}$ under hypothesis $\mathcal{H}_0$ and all positive-definite (PD) covariance matrices in $\mathbf{R}$ under hypothesis $\mathcal{H}_1$ such that $\mathbf{R}_1 = \mathbf{R}$, respectively. $\hat{\mathbf{R}}$ is the sample composite covariance matrix and $\hat{\mathbf{R}}_{ik}$ is the sample cross-covariance matrix of $\mathbf{R}$. And p is the probability density function (PDF) of $\mathbf{z}$. Note that, we assume that an experiment produces M *iid* (independently identically distribution) realizations of data matrix $\mathbf{X}$ [24], in this paper, we do EEMD with traffic data M times as the M independent copies, as shown in Fig. 5.

In [25], the distribution function of GLRT detector in the time domain under the null hypothesis was given as

$$\lambda | H_0 = \prod_{i=2}^{L} \prod_{n=0}^{N-1} Y_{in} \tag{16}$$

where $Y_{in} \sim Beta(\alpha_{in}, \beta_{in})$, and $\alpha_{in} = M - (i-1)N - n, n = 0, 1, \ldots, N-1$.

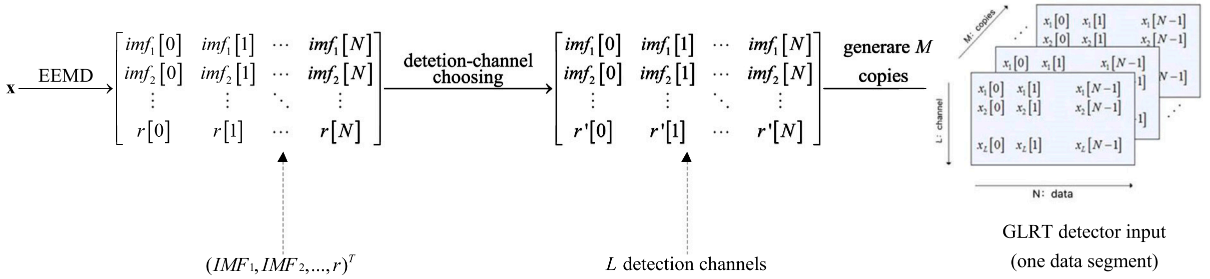


Fig. 5. Data preprocessing of multi-scale decomposition module.

3.3.2. Threshold and detection

Threshold T, as an important parameter to anomaly detection, represents the range of global decision without anomalies. We calculate T as following process:

1. Calculating λ with normal traffic using Eq. (15) and with Monte Carlo method, then a value-distribution chart of λ is got;
2. Given a predetermined false-alarm rate P_{FA}, find the corresponding λ according to the value-distribution chart. Then the threshold T is determined.

where, shown in Table 1, $P_{FA} = \frac{FP}{FP+TN}$.

With the probability distribution function that is derived from Eq. (16), we can calculate the exceedance fractions, which may be used to determine the **thresholds** T that correspond to a pre-defined P_{FA}.

Based on Eqs. (13) and (15), the detection function $\lambda(\mathbf{X}_{1:L}^{(1:t)})$ the detection decision rule can be formulated as

$$\begin{cases} \lambda(\mathbf{X}_{1:L}^{(1:t)}) > T & \text{declare } \mathcal{H}_1 \\ \lambda(\mathbf{X}_{1:L}^{(1:t)}) \le T & \text{declare } \mathcal{H}_0 \end{cases} \tag{17}$$

4. Experiments

4.1. Performance evaluation indicator

The Receiver operating characteristic (ROC) curve is a comprehensive indicator that reflects the continuous variable of sensitivity and specificity, is commonly used to measure the efficacy of anomaly detection methods [39]. The abscissa of the ROC curve is False positive rate (FPR) and the ordinate is True Positive rate (TPR), which are corresponding to false alarm probability P_{FA} and detection probability P_D. A good anomaly detection method can get the higher P_D with a lower P_{FA}. In other words, the closer the ROC curve is to the upper left corner, the better the performance of the anomaly detection method.

4.2. Datasets descriptions

Our experiments of anomaly detection focus on four datasets, including ISCX-IDS [40], MAWILab [41], and two datasets collected from campus networks in Chongqing higher education mega center (CHEMC) and an ISP company in China in 2017, respectively.

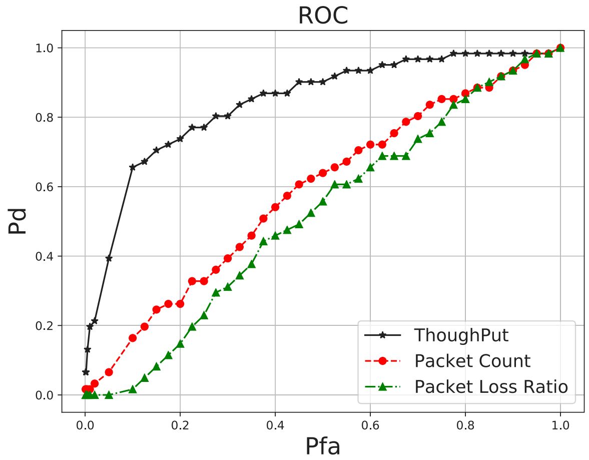


Fig. 6. Performance of different features.

ISCX-IDS is a widely-used simulation dataset generating from a systematic approach proposed in [40], recorded with 4 kinds of attack scenarios.

MAWILab is a real-world database that assists researchers to evaluate their traffic anomaly detection methods. The labels are obtained using an advanced graph-based methodology that compares and combines different and independent anomaly detectors. The data set is daily updated to include new traffic from upcoming applications and anomalies [41].

CHEMC dataset records the network traffic from 2014.12.22 12:01:12 to 2015.1.7 07:35:14. The anomaly here is the disconnection for an hour every night from 10:30 to 11:30.

ISP dataset anomalies were manually labeled by ISP experts based on actual conditions and data analysis and the types of anomalies are unknown.

All of our datasets were collected with an Ethernet tap, which could transmit the mirrored traffic without any processing overhead or disruption. The datasets are consisting of 49 columns, which represent 49 various features of the traffic while each column represents a time series with a specific feature. Summary of the features are shown in Table 2.

Table 2
Features of traffic data.

	Time	DownThroughout (Byte)	UpThroughout (Byte)	ClientJitter (ms)	ServerJitter (ms)	⋯	ClientTimeDelay (ms)	SeverTimeDelay (ms)	ServerPacketLossRatio (%)	ClientPacketLossRatio (%)
1	2010/6/14 12:00:00	3228	178	72	475	⋯	0	253	0	0
2	2010/6/14 12:00:01	4371	218	95	109	⋯	0	97	0	0
…						…				
n	2010/6/14 18:59:59	5686	263	6	68	⋯	1	24	0	0

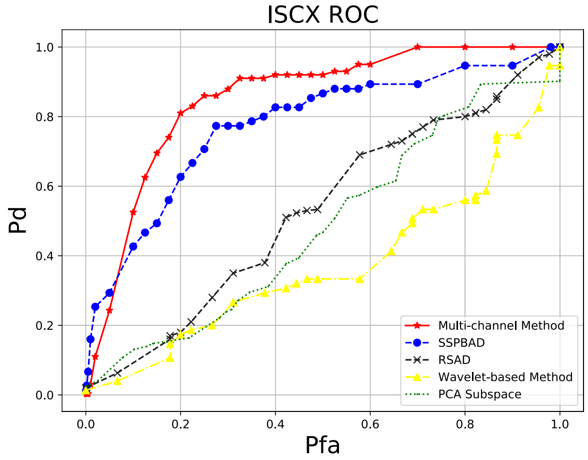


Fig. 7. Performance of ISCX-IDS on June 15th, 2010.

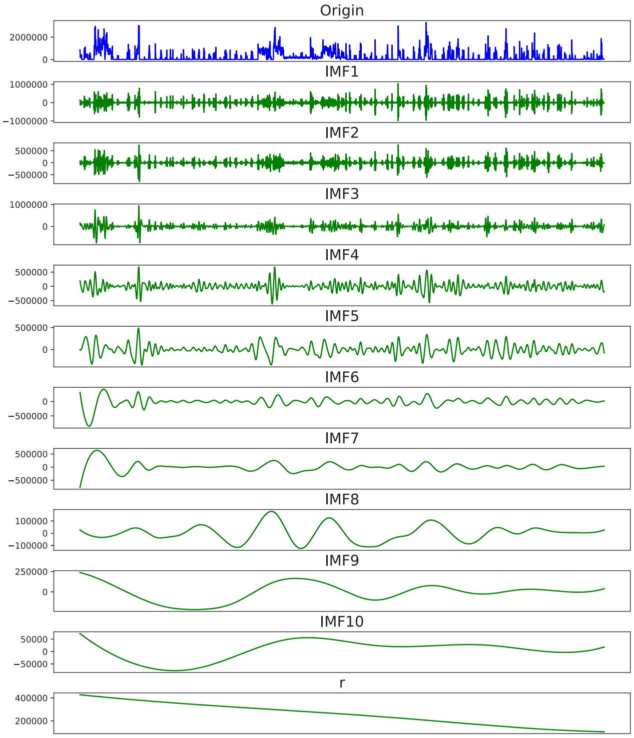


Fig. 8. The EEMD decomposition results on ISCX-IDS.

4.3. Feature selection

In order to obtain the feature with the highest correlation with the anomaly, we take the 49-dimensional traffic data vector as input, whether there is an anomaly as the label, and take the number of times the feature is selected as the GBDT split node as the feature importance. Here, the main parameters of the GBDT are: the maximum number of leaves is 30, the maximum number of iterations is 100, the learning rate is 0.05.

The result is as shown in Table 3. We can see that up throughput and down throughput are of the highest importance, in the information fusion step, we compress the two columns of features into one dimension, so the following anomaly detection is carried out with the compressed one-dimensional data.

Table 3
Features importance.

Features		Importance	
Throughput	Up Down	69 48	117
Packet count	Up Down	42 41	83
Jitter	Client Server	42 34	76
Time delay	Client Server	6 40	46
Packet loss ratio	Clint Server	8 12	20

Table 4
The component average period summary.

	Average period (s)	
	Normal traffic	All traffic
IMF_1	3.46	3.32
IMF_2	7.35	7.09
IMF_3	14.88	14.43
IMF_4	29.88	30.38
IMF_5	56.25	60.00
IMF_6	102.86	128.57
IMF_7	218.18	240.00
IMF_8	450.00	800.00
IMF_9	900.00	1200.00
IMF_{10}	1800.00	1800.00
r	7200.00	7200.00

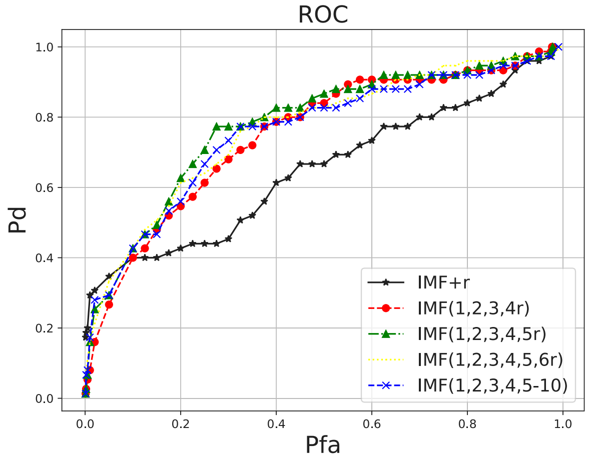


Fig. 9. Influence of difference channel selection scheme.

4.4. Experiments on ISCX-IDS

In this subsection, experiments were carried out with the ISCX-IDS dataset, which was recorded from 12:00:00 to 14:59:59 in June 15th, 2010.

First, do feature selections and fusion for the traffic vector, and the importance of the feature is shown in Table 4. Then, the traffic data is decomposed into 10 IMFs and a residual r by EEMD, which are shown as Fig. 8. We can see from Fig. 8 that no prominent abnormalities could be observed from either original data and the multi-scale components decomposed from original data. Thus further detection procedures on decomposed data must be carried out. Detection procedures can be divided into two parts, threshold calculation and anomaly detection. As mentioned in Section 3.2, we used traffic data from 12:00:00 to 15:00:00 as threshold calculation period (i.e. no abnormalities appeared) and traffic data from 17:00:00 to 18:00:00 as anomaly-detection period with the dataset.

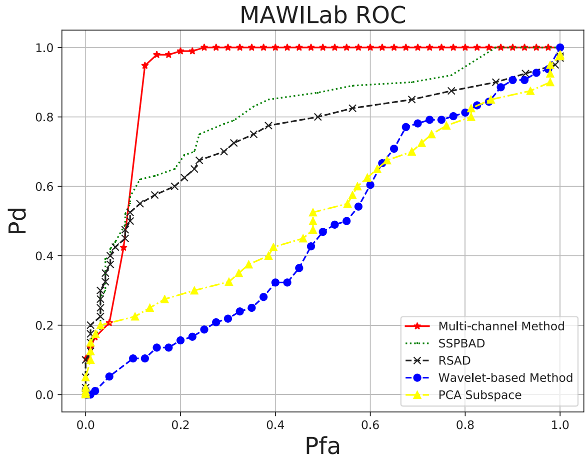


Fig. 10. The results based on the data of MAWILab.

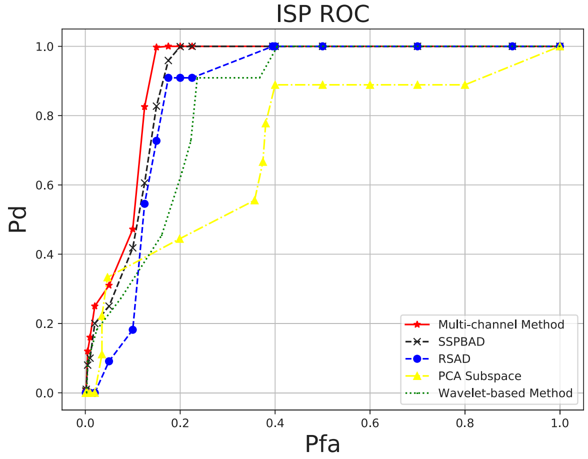


Fig. 11. The results based on the data of ISP company.

As shown in Table 4, the original data was decomposed into 11 components including residue r, Each component has a different period, that is, the frequency is also different. Considered with the channel-chosen principle described in Section 2, IMF_1 to IMF_4 are regarded as the individual detection channel respectively, while other components would be regarded as integrated channel. Thus, the total amount of the detection channel $L = 5$. We also set $N = 30$ and $M = 50$ in our experiments.

Different channel selection schemes are compared using ISCX-IDS datasets. The ROC curves obtained with different detection channels are shown in Fig. 9, where the channel-chosen principle described before is expressed as IMF(1,2,3,4,5r), which contains 5 detection channels. IMF_1 to IMF_4 are regarded as the individual detection channel respectively, while other components (from IMF_5 to r) would be regarded as integrated channel. It is found that the anomaly detection results are affected by the channel integrity.

We can see from Fig. 9 that the scheme of IMF+r, which means each component is regarded as an individual channel, performs worst. Other four channel selection schemes are approximate, while our 5-channel scheme performs better.

There are several reference network traffic anomaly detection methods: wavelet-based [27], PCA-based [26], SSPBAD [22], RSAD [22] were considered in this part, as the performance comparisons. The ROC curves obtained by each method are shown in Fig. 7, the performance of our anomaly detection method is better than other three detection methods on the same dataset. And The ROC of different features by our method is shown in Fig. 6.

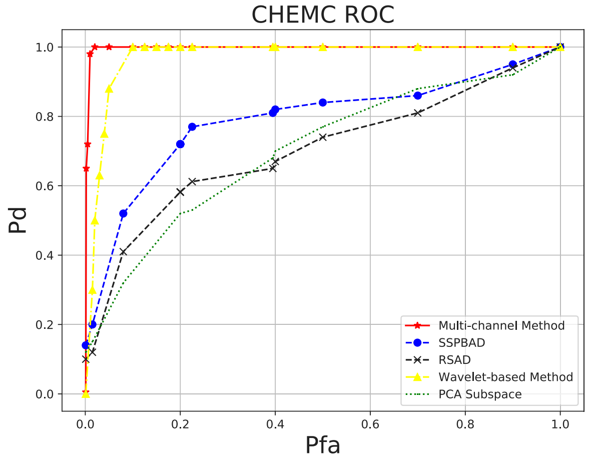


Fig. 12. The results based on the data of Chongqing higher education mega center.

Table 5
AUC of methods on datasets.

Area Under Curve(AUC)				
	MAWILab	USCX-IDS	ISP	CHEMC
Multi-channel method	0.9212	0.8522	0.9199	0.9965
SSPBAD	0.8092	0.7760	0.9030	0.7887
RSAD	0.7492	0.5282	0.8664	0.7013
Wavelet	0.4809	0.3733	0.7181	0.9718
PCA Subspace	0.5340	0.4860	0.8430	0.7015

4.5. Experiments on MAWILab

In this subsection, experiments were carried out with the MAWILab, we selected the network traffic data on January 1, 2019. This data records a total of 900 s of traffic data from January 1, 2019 14:00:00 to January 1, 2019 14:15:00, with 53 554 628 packets (19 709.38 MB) and 165 anomalies.

The protocol breakdown of this data is shown in Table 6, it can be seen that there are many types of protocols involved. And the anomalies of this dataset and the network environment are more complicated than ISCX-IDS, therefore we believe that this dataset can better evaluate the detection performance of each anomaly detection method on real-world complex conditions.

After the same steps as the above dataset, we get the experimental results, as shown in Fig. 10, we can see that the method proposed in this paper performs significantly better on this data set than other methods. Although it performs worse than other methods when the P_{FA} is low, when the P_{FA} is only 20%, the P_D reaches almost 100%. The P_D of other methods is relatively low at a P_{FA} of 20%, which is far behind the method proposed in this paper.

4.6. Experiments on CHEMC and ISP datasets

With other two datasets, collected from campus networks in CHEMC and an ISP company, the experiments were carried out following the same steps with ISCX-IDS, the topology of these tests is as shown in Fig. 1, without NAT server. The results are shown in Figs. 11 and 12.

It can be learnt that our method still performs better than other three traditional methods. With the data of CHEMC, while P_{FA} equals to 0.02, the P_D of our method reaches 100%. With the same condition, the P_D of other methods are 5% and 35%., respectively. With the data collected from an ISP, while P_{FA} equals to 15%, the P_D of our method is more than 90%. With the same condition, the P_D of other methods are both at 45%.

Table 6
MAWILab data protocol breakdown.

Protocol			Packets	Bytes	Bytes/pkt
IP	TCP	HTTP	4 301 821 (8.03%)	3 175 621 967 (15.37%)	738.20
		HTTPS	8 372 949 (15.63%)	9 764 098 322 (47.25%)	1166.15
		SMTP	75 886 (0.14%)	25 602 604 (0.12%)	337.38
		FTP	18 227 (0.03%)	1 077 474 (0.01%)	59.11
		SSH	813 712 (1.52%)	138 752 570 (0.67%)	170.52
		DNS	14 198 (0.03%)	960 785 (0.00%)	67.67
		BGP	2548 (0.00%)	357 243 (0.00%)	140.21
		Other	10 037 411 (18.74%)	1 623 956 197 (7.86%)	161.79
		Total	23 636 752 (44.14%)	14 730 427 162 (71.28%)	623.20
	UDP	DNS	291 379 (0.54%)	44 815 979 (0.22%)	153.81
		Other	2 749 884 (5.13%)	1 403 317 888 (6.79%)	510.32
		Total	3 041 283 (5.68%)	1 448 138 731 (7.01%)	476.16
	Other	ICMP	22 040 489 (41.16%)	1 343 057 815 (6.50%)	60.94
		GRE	3 253 505 (6.08%)	1 904 946 551 (9.22%)	585.51
		IPSEC	4340 (0.01%)	3 324 404 (0.02%)	765.99
		FRAG	333 (0.00%)	324 083 (0.00%)	973.22
	Total		51 976 669 (97.05%)	19 429 977 196 (94.02%)	373.82
IP6	Total		1 577 957 (2.95%)	1 236 809 155 (5.98%)	783.80
Total			53 554 628 (100.00%)	20 666 786 471 (100.00%)	385.90

4.7. Results analysis

Since different data sets are acquired under different conditions and topologies, the detection performances of these methods are changed with different datasets, as shown in Table 5, Figs. 7, 10–12. We believe the quantity of samples, the network topology structure, the types of anomalies and the methodologies of these detection methods caused the changes of detection performance of methods.

In detail, the network topology structure of the Mawilab dataset is very complicated [41], there are a large number of nodes in the network, the method proposed in this paper will be significantly better than other methods on this dataset because our method can capture the sufficient spatial correlation information from the data, as shown in Fig. 10. However, in the ISCX-IDS dataset, there are only dozens of nodes in the network [40], thus our method did not obtain a prominent detection performance compared with other methods, as shown in Fig. 7. Considering the number of samples in datasets, our method obtained a better performance for the large number of samples, while PCA method obtained a worse performance, as shown in Fig. 11. And Considering the types of anomalies, the Mawilab dataset contains the largest number of various types of anomalies, while the CHEMC dataset contains only one simple type of network anomaly, thus the performance of methods changed significant with the 2 datasets, as shown in Figs. 10 and 12.

According to the experimental results and theoretical analysis, we can learn that, for our method, when the network topology is complicated, which means our method can capture enough spatial information; the data samples are sufficient, which means more temporally information; and the dataset contains anomalies that can influence the network traffic features of multiple nodes in the network like DDoS attacks, flooding attacks etc., our method can fully consider the internal spatio-temporal correlations, and the ideal detection performance will be obtained. In other words, our method can achieve an ideal detection accuracy for detecting complicated network anomalies under the condition of large-scale data in complicated network topology structure, which also meets the needs of current situation.

Overall, it can be seen that even though each method has unstable performance, our method performs the best on all the datasets as shown in Table 5. This shows that the method proposed in this paper has a good anomaly detection performance regardless of the amount of data sample and the condition of anomalies contained in datasets. The effectiveness of the method proposed in this paper is proved by the experiment results.

5. Conclusion

This paper proposes a multi-channel anomaly detection method based on signal detection theory and multi-scale decomposition. Our method firstly combines EEMD and multi-channel GLRT to perform anomaly detection. Compared with traditional multi-scale detection methods, this method fully considers the internal frequency–time correlations within multiple scales of traffic data. It can be shown with the experiment results that our method can perform better with wider applicability than other traditional multi-scale anomaly detection methods.

Our proposed method still needs to be improved in several aspects. Firstly, more theoretical channel selection method that is more suitable for adaptive multi-channel detection should be researched in the future. Secondly, only preliminary experiments were carried out under very limited features. More complex scenarios/datasets will be considered and more comparisons with other detection algorithms should be carried out. Thirdly, artificial intelligence could be involved into our proposed scheme, which might be make our scheme more practical.

CRediT authorship contribution statement

Lisheng Huang: Conceptualization, Formal analysis, Validation. **Jinye Ran:** Methodology, Writing - original draft, Software. **Wenyong Wang:** Reviewing, Supervision, Investigation. **Tan Yang:** Validation, Visualization. **Yu Xiang:** Conceptualization, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China, No. 2018YFB08040505; Sichuan Science and Technology Program, China, No. 18ZDYF2573; The authors wish to thank the reviewers and editors for their helpful comments.

References

[1] K.-K.R. Choo, The cyber threat landscape: Challenges and future research directions, Comput. Secur. 30 (8) (2011) 719–731.

[2] M. Ahmed, A.N. Mahmood, J. Hu, A survey of network anomaly detection techniques, J. Netw. Comput. Appl. 60 (2016) 19–31.

[3] F.Y. Edgeworth, Xli. on discordant observations, London, Edinburgh, Dublin Phil. Mag. J. Sci. 23 (143) (1887) 364–375.

[4] F.J. Anscombe, Rejection of outliers, Technometrics 2 (2) (1960) 123–146.

[5] E. Eskin, Anomaly detection over noisy data using learned probability distributions, 2000.

[6] R. Laxhammar, G. Falkman, Online learning and sequential anomaly detection in trajectories, IEEE Trans. Pattern Anal. Mach. Intell. 36 (6) (2013) 1158–1173.

[7] A. Siffer, P.-A. Fouque, A. Termier, C. Largouet, Anomaly detection in streams with extreme value theory, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 1067–1075.

[8] M. Desforges, P. Jacob, J. Cooper, Applications of probability density estimation to the detection of abnormal conditions in engineering, Proc. Inst. Mech. Eng., Part C 212 (8) (1998) 687–703.

[9] F. Simmross-Wattenberg, J.I. Asensio-Perez, P. Casaseca-De-La-Higuera, M. Martin-Fernandez, I.A. Dimitriadis, C. Alberola-Lopez, Anomaly detection in network traffic based on statistical inference and\alpha-stable modeling, IEEE Trans. Dependable Secure Comput. 8 (4) (2011) 494–509.

[10] G. Poojitha, K.N. Kumar, P.J. Reddy, Intrusion detection using artificial neural network, in: 2010 Second International Conference on Computing, Communication and Networking Technologies, IEEE, 2010, pp. 1–7.

[11] I. Syarif, A. Prugel-Bennett, G. Wills, Unsupervised clustering approach for network anomaly detection, in: International Conference on Networked Digital Technologies, Springer, 2012, pp. 135–145.

[12] S. Naseer, Y. Saleem, S. Khalid, M.K. Bashir, J. Han, M.M. Iqbal, K. Han, Enhanced network anomaly detection based on deep neural networks, IEEE Access 6 (2018) 48231–48246.

[13] K. Demertzis, L. Iliadis, A hybrid network anomaly and intrusion detection approach based on evolving spiking neural network classification, in: International Conference on E-Democracy, Springer, 2013, pp. 11–23.

[14] I. Balabine, A. Velednitsky, Method and system for confident anomaly detection in computer network traffic, 2017, US Patent 9, 843, 488.

[15] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. (9) (2008) 1263–1284.

[16] B.A. Tama, M. Comuzzi, K. Rhee, Tse-ids: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system, IEEE Access 7 (2019) 94497–94507.

[17] B.A. Tama, L. Nkenyereye, S.M.R. Islam, K. Kwak, An enhanced anomaly detection in web traffic using a stack of classifier ensemble, IEEE Access 8 (2020) 24120–24134.

[18] Y. Zhong, W. Chen, Z. Wang, Y. Chen, K. Li, Helad: A novel network anomaly detection model based on heterogeneous ensemble learning, Comput. Netw. 169 (2020) 107049.

[19] M. Thottan, C. Ji, Anomaly detection in IP networks, IEEE Trans. Signal Process. 51 (8) (2003) 2191–2204.

[20] D. Brauckhoff, K. Salamatian, M. May, Applying PCA for traffic anomaly detection: Problems and solutions, in: IEEE Infocom 2009, IEEE, 2009, pp. 2866–2870.

[21] H. Ren, M. Liu, X. Liao, L. Liang, Z. Ye, Z. Li, Anomaly detection in time series based on interval sets, IEEJ Trans. Electr. Electron. Eng. 13 (5) (2018) 757–762.

[22] M.F. Kaloorazi, R.C. de Lamare, Anomaly detection in IP networks based on randomized subspace methods, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 4222–4226.

[23] D. Jiang, Z. Xu, P. Zhang, T. Zhu, A transform domain-based anomaly detection approach to network-wide traffic, J. Netw. Comput. Appl. 40 (2014) 292–306.

[24] D. Ramírez, J. Vía, I. Santamaría, L.L. Scharf, Detection of spatially correlated Gaussian time series, IEEE Trans. Signal Process. 58 (10) (2010) 5006–5015.

[25] N. Klausner, M.R. Azimi-Sadjadi, L.L. Scharf, Detection of spatially correlated time series from a network of sensor arrays, IEEE Trans. Signal Process. 62 (6) (2014) 1396–1407.

[26] D. Jiang, C. Yao, Z. Xu, W. Qin, Multi-scale anomaly detection for high-speed network traffic, Trans. Emerg. Telecommun. Technol. 26 (3) (2015) 308–317.

[27] R. Fontugne, P. Abry, K. Fukuda, P. Borgnat, J. Mazel, H. Wendt, D. Veitch, Random projection and multiscale wavelet leader based anomaly detection and address identification in internet traffic, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 5530–5534.

[28] I.C. Paschalidis, G. Smaragdakis, Spatio-temporal network anomaly detection by assessing deviations of empirical measures, IEEE/ACM Trans. Netw. 17 (3) (2009) 685–697.

[29] I. Nevat, D.M. Divakaran, S.G. Nagarajan, P. Zhang, L. Su, L.L. Ko, V.L. Thing, Anomaly detection and attribution in networks with temporally correlated traffic, IEEE/ACM Trans. Netw. 26 (1) (2018) 131–144.

[30] A. Soule, K. Salamatian, N. Taft, Combining filtering and statistical methods for anomaly detection, in: Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement, USENIX Association, 2005, p. 31.

[31] Z. Wu, N.E. Huang, Ensemble empirical mode decomposition: a noise-assisted data analysis method, Adv. Adapt. Data Anal. 1 (01) (2009) 1–41.

[32] J. Yang, P. Li, Y. Yang, D. Xu, An improved EMD method for modal identification and a combined static-dynamic method for damage detection, J. Sound Vib. 420 (2018) 242–260.

[33] Y. Xiang, X. Wang, L. He, W. Wang, W. Moran, Spatial-temporal analysis of environmental data of north Beijing district using Hilbert-huang transform, PLoS One 11 (12) (2016) e0167662.

[34] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232.

[35] X. Yuan, M. Abouelenien, A multi-class boosting method for learning from imbalanced data., IJGCRSIS 4 (1) (2015) 13–29.

[36] A. Hervé, L.J. Williams, Principal component analysis, Wiley Interdiscip. Rev. Comput. Stat. 2 (4) (2010) 433–459.

[37] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chem. Intell. Lab. Syst. 2 (1–3) (1987) 37–52.

[38] N.E. Huang, An adaptive data analysis method for nonlinear and nonstationary time series: the empirical mode decomposition and Hilbert spectral analysis, in: Wavelet Analysis and Applications, Springer, 2006, pp. 363–376.

[39] J. Davis, M. Goadrich, The relationship between precision-recall and ROC curves., in: Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 233–240.

[40] A. Shiravi, H. Shiravi, M. Tavallaee, A.A. Ghorbani, Toward developing a systematic approach to generate benchmark datasets for intrusion detection, Comput. Secur. 31 (3) (2012) 357–374.

[41] R. Fontugne, P. Borgnat, P. Abry, K. Fukuda, Mawilab: combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking, in: Proceedings of the 6th International Conference, 2010, pp. 1–12.

Lisheng Huang was born in Chongqing, China, in 1975. He received the Ph.D. degree from University of Electronic Science and Technology of china (UESTC), in 2011. He is currently an associate research fellow of UESTC. His research interests include network measurement/management/security.

Jinye Ran received the bachelor's degree in digital media technology from Shandong University in 2017, China, and received the master's degree in Computer Science from University of Electronic Science and Technology of China (UESTC) in 2020 under the direction of Dr. L. Huang and Dr. Y. Xiang. His research interests include network security and network anomaly detection.

Wenyong Wang was born in 1967. He is now a Professor of Computer Science at University of Electronic Science and Technology of China (UESTC). He holds a B.E. in Computer Science from Beijing University of Aeronautics and Astronautics, Beijing, China, an M.E. in Computer Science and a Ph.D. in Communications Engineering from UESTC.

His research interests include network architecture, Internet measurement and performance management, and wireless sensor networks and their applications. He is a member of IEEE, a senior member of CCF, and a member of the council of Internet Society of China.

Tan Yang received her Ph.D degree from Beijing University of Posts and Telecommunications in 2010. She is now a associate professor in Beijing University of Posts and Telecommunications. Her research interests are network performance evaluation and mobile Internet.

Yu Xiang received M.S. and Ph.D. degrees from University Electronic Science and Technology of China (UESTC), Chengdu, Sichuan, China, in 1998 and 2003, respectively.

He joined UESTC in 2003 and became an associate professor in 2006. From 2014–2015, he was a visiting scholar in the University of Melbourne, Australia. His current research interests include network anomaly detection, wireless sensor networks, IOT and ITS.