Air Quality Data Clustering using EPLS Method

Yunliang Chen, Lizhe Wang, Fangyuan Li, Bo Du,
Kim-Kwang Raymond Choo, Houcine Hassan, Wenjian Qin

PII: S1566-2535(16)30196-8
DOI: 10.1016/j.inffus.2016.11.015
Reference: INFFUS 826

To appear in: *Information Fusion*

Received date: 16 September 2016
Revised date: 11 November 2016
Accepted date: 29 November 2016

Please cite this article as: Yunliang Chen, Lizhe Wang, Fangyuan Li, Bo Du, Kim-Kwang Raymond Choo, Houcine Hassan, Wenjian Qin, Air Quality Data Clustering using EPLS Method, *Information Fusion* (2016), doi: 10.1016/j.inffus.2016.11.015

Highlights

- An approach EPLS is proposed for air quality data fusion and clustering.
- EPLS preserves the most valuable features which are adaptive to other measures and clustering approaches.
- EPLS-based clustering algorithm can easily handle large-volumes of data.
- EPLS can be efficiently suitable for air quality clustering problem.

Air Quality Data Clustering using EPLS Method☆

Yunliang Chen[a,e,1], Lizhe Wang[a,*], Fangyuan Li[a,**], Bo Du[a], Kim-Kwang Raymond Choo[b], Houcine Hassan[c], Wenjian Qin[d]

[a] *China University of Geosciences, Wuhan, China*
[b] *University of Texas at San Antonio, USA and University of South Australia, Australia*
[c] *Polytechnic University of Valencia, Spain*
[d] *Shenzhen Institute of Advanced Technology, Chinese Academy of SciencesUniversity of Chinese Academy of Sciences*
[e] *Hubei Key Laboratory of Intelligent Geo-information Processing, China University of Geosciences, Wuhan 430074, China*

Abstract

Nowadays air quality data can be easily accumulated by sensors around the world. Analysis on air quality data is very useful for society decision. Among five major air pollutants which are calculated for AQI (Air Quality Index), $PM_{2.5}$ data is the most concerned by the people. $PM_{2.5}$ data is also cross-impacted with the other factors in the air and which has properties of non-linear non-stationary including high noise level and outlier. Traditional methods cannot solve the problem of $PM_{2.5}$ data clustering very well because of their inherent characteristics. In this paper, a novel model-based feature extraction method is proposed to address this issue. The EPLS model includes 1) Mode Decomposition, in which EEMD algorithm is applied to the aggregation dataset; 2) Dimension Reduction, which is carried out for a more significant set of vectors; 3) Least Squares Projection, in which all testing data are projected to the obtained vectors. Synthetic dataset and air quality dataset are applied to different clustering methods and similarity measures. Experimental results demonstrate

☆Fully documented templates are available in the elsarticle package on CTAN.

*Corresponding author at: Department of Computer Science, China University of Geosciences, Wuhan, China.

**Corresponding author at: Department of Computer Science, China University of Geosciences, Wuhan, China.

Email addresses: lizhewang@icloud.com (Lizhe Wang), lfy_cug@hotmail.com (Fangyuan Li)

[1]Email address: Cyl_king@hotmail.com.

that EPLS is efficient in dealing with high noise level and outlier air quality clustering problems, and which can also be adapted to various clustering techniques and distance measures.

Keywords: Air Quality, $PM_{2.5}$, Clustering, EEMD, PCA

1. Introduction

In recent years, an increasing number of sensor devices have generated a large amount of temporal data which can be treated as time series data. These time series can be measured and analyzed across the scientific disciplines, including human beats in medicine, cosmic rays in astrophysics, rates of inflation in economics, and air temperatures in climate science[1] etc. Extracting numerical features from time series data would have a huge influence for human decision, such as revealing human interpretable characteristics of the human activity data [2], data forecasting for social behaviors as well as clustering and classification [3, 4, 5].

According to the definition of AQI, ambient air pollutants in China are concentrations of particulate matter ($PM_{2.5}$ and PM_{10}), SO_2, CO, NO_2, O_3. These time series are nonstationary and seasonality with high level of noise and outlier. Nonstationary means that the statistic properties change with time. Many studies have analyzed air quality time series in the stationary framework, however, this assumption is invalid. Air quality time series should be pre-processing to deal with the negative impact caused by noise and outlier. Traditional whole time series clustering and statistic feature extraction are unable to overcome the patterns of noise and outlier. In order to solve these problem, a novel algorithm needed to be proposed.

Among all techniques applied to analyzing time series data, clustering is the most widely used one without costly human supervision or time-consuming annotation of data [6]. Clusters are formed by grouping objects which have maximum similarity in an unlabelled dataset. In detail, clustering is applied on exploratory data for summary generation and acts as a pre-processing step or

subroutine for other tasks [7, 8, 9]. It can be concluded that time series clustering can be classified into three main branches [10]:(1) whole time series clustering [11]; (2) subsequence clustering [7]; (3) time point clustering [12]. As for whole time series clustering, there are three different categories, namely shape-based [6, 13, 14], feature-based [2, 10] and model-based [15, 16]. In the shape-based approach, two time-series are matched by shapes and usually employ conventional clustering methods[17]. As for feature-based approach, the raw time series are converted into a feature vector for clustering methods[18]. Model-based approach is transformed from raw time series into model parameters (a parametric model for each time series), and then a suitable model distance and a suitable clustering algorithm can be chosen.

According to nonstationary time series, feature-based representations of time series are universal. Some factors such as mean, standard deviation, skewness and kurtosis are used as features by Nanopoulos [19]. These features are statistical values for stationary time series with low dimensions. Usue et al. introduced ten features that contains measures of dimension, shift, correlation, seasonality, trend, noise, outliers, autocorrelation, skewness and kurtosis to represent time series [10], and these features form a characteristic vector for clustering[20]. Valchos et al. used periodic features to obtain via the direct Fourier Decomposition for clustering of MSN query log and electrocardiography time series data [21]. Duncan et al. describes a new time-frequency feature extraction method, which is based on Empirical Mode Decomposition (EMD) [2]. Generally, these features are classified into three categories: (1) Time domain features. These techniques usually involve extracting statistic (e.g., variance, mean, spread, Gaussianity), or some information theoretic and entropy/complexity measures(e.g., automutual information, Approximate Entropy, Lempelziv complexity). (2) Frequency domain features [22]. These methods are mostly underpinned by the discrete Fourier or Wavelet transformation of data. (3) Time-frequency domain features. These features can be obtained from Hilbert-Huang transformation, Wignerville distribution, Cohen time-frequency distribution.

After features extraction, a similarity measure need to be selected, and Eu-

clidean distance can lead to better clustering results as an useful method [11]. But Euclidean distance measure is not a generally method, for some datasets Elastic measure including DTW and edit distance can achieve higher performance. Is there a better distance measure method for general dataset? In the same time, the choice of features representation is also a problem [23]. For a specific dataset, the selection of feature vector and suitable distance measure is a very challenging task[24], and these two measures can extremely affect the results of time series clustering.

In this paper, EPLS, a novel algorithm, is proposed for feature based time series clustering. EPLS algorithm is based on Ensemble Empirical Mode Decomposition (EEMD), Principal Component Analysis (PCA) and Least Square method (LS), which transfers the raw time series into a new feature time series within ten dimensions. EEMD is expert in dealing with large-scale non-stationary time series, and which can reveal the inner time-frequency patterns. This is prominent for air quality time series clustering. Usually the raw time series is subjected to high noise and outlier levels, meanwhile the high-dimensional feature was affected by some distance measures. Therefore, it is difficult to achieve high performance for directing clustering. EPLS attempts to provide the features of time series which reserves the most valuable feature vector. This feature vector is distance measure independent for its low-dimensional characteristic, and clustering method independent for low noise and outliers levels.

The paper is organized as follows. In Section 2 the related work is discussed including the issue of using other feature extraction techniques. Then an introduction about the new algorithm EPLS, a combination of EEMD, PCA and LS is detailed in Section 3. Experiment settings are described in Section 4. Section 5 gives the experimental results. We conclude the contribution of our work in Section 6.

2. Related Work

Air quality time series study has significant influence on social coherence and national economic. Many studies have developed to analyze air quality time series [25]. Mohammad [26] detects the change in climate time series based on Bounded-Variation Clustering. Kassomenos et al. [27] used principal component analysis and regression analysis to quantify the contribution of both combustion and non-combustion sources to the PM_{10} and $PM_{2.5}$ levels in Athens. Many researchers have adopted ANNs to predict $PM_{2.5}$ or other air pollutant factors. Voukantsis et al. and Qingping Zhou et al. [28, 29] combined ANNs with other algorithms like principle component analysis, correlation coefficient analysis, to predict daily $PM_{2.5}$ concentration. However, the original air quality time series is always nonlinear with sharp transition and non-stationary with different frequency characteristics [30]. These approaches are various some based on clustering and some based on forecasting. In this work, the target is to analyze air quality time series upon clustering method. Numerous techniques for time series clustering have been proposed. The most prevalent use of the proposed techniques are based on feature extraction, both time and frequency domain. Two main challenges of feature-based time series clustering are typical: (1) selecting an appropriate feature vector of the time series, and (2) choosing a suitable similarity measures or distance between time series [31].

The operations to quantify time series properties are various. The basic statistics values including location, trend, Gaussianity, outlier and noise are usually computed [32]. The properties of stationary, linear correlations and information theoretic measure are also universal standards. The wavelet methods, properties of network derived from time series are the model fits features [33]. All of these different feature extraction techniques are all transferring time series $X = (x_1, x_2, ..., x_n)$ into some real numbers [1]. These real numbers form the feature vector, but there are hundreds of features. So it is a question to select the best suitable features. Otherwise, the human related time series are usually characterized with high level of noise and outlier [34]. The statistic val-

ues based on time domain can not deal with such kind of time series, or the straightforward clustering is incapable of performing well.

Experiments suggest that not all of these distance measures are appropriate for all time series [31]. This is probably because of the various characteristics of time series, which results in some distance measures more suitable than others. Time series clustering relies on distance measure to a high extent. The theoretical study of time series similarity search is proposed by Argawal et al [35]. The Dynamic Time Warping (DTW), Euclidean distance (ED), Edit Distance for Real Sequence (EDR), HMM-based distance, Fourier coefficient based similarity measure, and Longest Common Sub-Sequence (LCSS) are the most popular distance measures that are used for time series data. ED is one of the simplest measures applied in data mining. Although ED is widely used in many fields, some research points out that ED is not an appropriate measure for time series analysis [31]. First, ED is only suitable for dealing with time series with equal length; and that ED is highly susceptible to noise and outliers. DTW is the most popular, which can deal with transformations such as local warping and shifting, furthermore, it is capable to compare time series with different length [31]. However, its complexity is $O(mn)$, where m and n are the length of two time series, respectively. As for EDR, it has shown rather positive results in previous work [36]. In some cases, EDR becomes more flexible and capable when dealing with noisy data or outliers. In a word, a distance measure can not apply in every field with better performance. Most studies pay attention to select appropriate distance measures for different time series, which is hard and complex. In addition, there are several studies transferring raw time series into a new time series. The new time series can be measured by simple distance measures like ED, and thus has lower level of noise and outliers.

3. Mode decomposition, Component analysis and Projection (EPLS)

This project aims at developing a domain-independent, accurate, and scalable feature extraction method. The target of this time-frequency method is to

find a base vector for a time series dataset. Then all the time series is mapped to the base vector, and a new set of time series can be obtained. The extracted features from time series is to serve as inputs for machine learning algorithms. EPLS as a way of dealing with features extraction, the outline of which is shown below. Firstly, all time series in a dataset are pooled into aggregation and EEMD algorithm is applied to this aggregation. Then a dimension reduction operation is carried out for a more significant set of vectors. Finally, all time series from a dataset are projected to the obtained vectors. There are three stages for implementing the EPLS algorithm including mode decomposition stage, dimension reduction and projection, as Fig.1 shows.

3.1. Mode Decomposition

Given dataset including n time series $X = (x_1, x_2, ..., x_n)$, where each time series in X is a vector of length m, that is, $\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{im})$. Hence, all the n time series is aggregated as follows

$$R = \sum_{i=1}^{n} x_i \tag{1}$$

R is the representation of a dataset. The input of EEMD algorithm is R time series. EEMD algorithm consists of sifting an ensemble of white noise-added signal and treats the mean as the final true result [37]. Compared with Empirical Mode Decomposition (EMD), EEMD algorithm adds white noise to provide a uniform reference frame in the time-frequency space; thus the added noise collates the portion of the signal of comparable scale in one Intrinsic Mode Function (IMF). The EEMD algorithm overcomes two problems: the end effect and the stoppage criteria from EMD algorithm [38]. EEMD signal decomposition technique is utilized to decompose R into num IMFs.

$$R = \sum_{j=1}^{num} imf_j \tag{2}$$

where subSeries $\mathbf{imf_j} = (imf_{j1}, imf_{j2}, imf_{j3}, ..., imf_{jm})$ corresponds to a time-frequency component. SubSeries $\mathbf{imf_1}, \mathbf{imf_2}, ..., \mathbf{imf_{num}}$ is relevant, and is or-

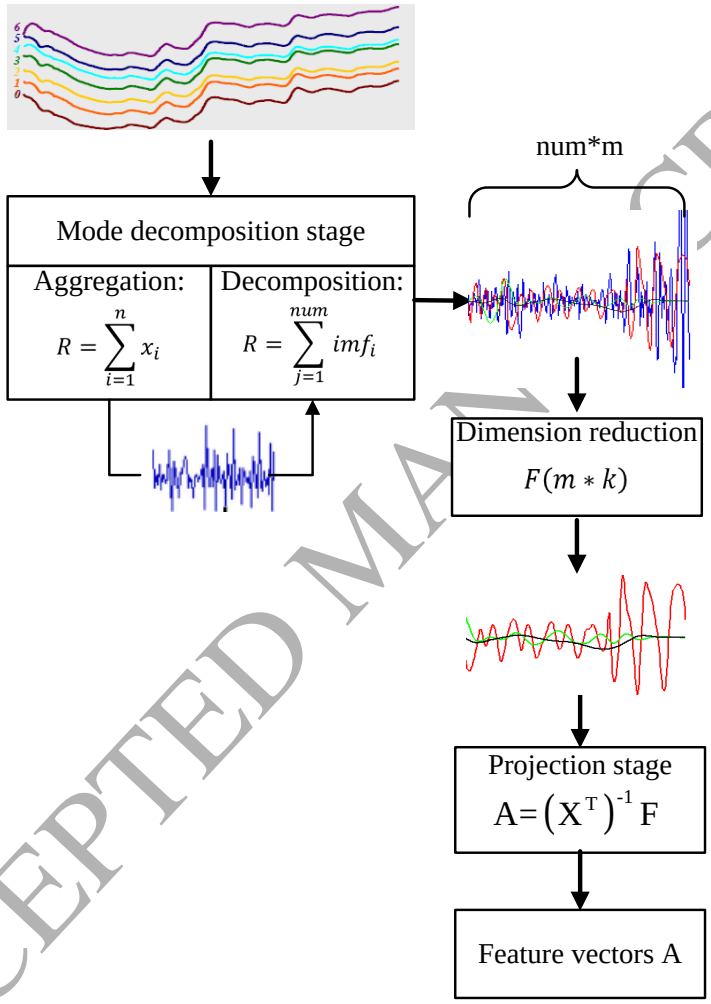

Figure 1: The procedure flow of EPLS

dered in descending order of frequency. In this process, there are no time-frequency components discarded.

3.2. Dimension reduction

SubSeries (2) have relative high dimension, and some components can be linear relevant. These feature will prevent projecting stage from getting better results. Principal Component Analysis(PCA) is a standard tool in modern data analysis for diverse fields (from neuroscience to computer graphics) [39]. PCA is a simple, non-parametric method for extracting relevant information from confusing data sets. In this process, PCA is applied in dimension reduction and obtaining totally new orthogonal vectors. The main idea of PCA is to project the $m-dimension$ features into $k-dimension$ features ($k < m$), and the $k-dimension$ features are orthogonal vectors.

Applying PCA technique in subSeries (2) in following steps: (1) To compute the *mean* of R of every column, then compute the new value D(m*num) which represents the original time series minus mean value

$$D(m*num) = \sum_{j=1}^{num} (imf_j - mean_j)^T \tag{3}$$

(2) To compute covariance matrix; (3) To obtain eigenvalues and eigenvectors from covariance matrix; (4) To order eigenvalues in descending order, and pick up the biggest k eigenvalues to construct eigenvector matrix-$EigenVectors(num*k)$; (5) To project sample vector R into picked eigenvector matrix

$$F(m*k) = D(m*num) \cdot EigenVectors(num*k) \tag{4}$$

So far, the original sample is changed from $num-dimension$ to $k-dimension$ and the dimension reduction is realized. $F(m*k)$ is the result after applying PCA algorithm .This procedure ensures that only the components which correspond to the most typical and irrelevant time-frequency patterns of the aggregate are selected. Meanwhile, PCA technique removes the noise and redundancy. Additionally, vector $F(m*k)$ is standardized, i.e.

$$\hat{F}_j = \frac{2 * (F_j - F_{jmin})}{F_{jmax} - F_{jmim}} - 1, j = 1, 2, ..., k. \quad (5)$$

where F_{jmin} is the minimum value of $\mathbf{F_j}$, and F_{jmax} is the maximum value, $\hat{F}_j \in [-1, 1]$. The orthogonality, standardization and the linear independence ensure that, $\hat{F}$ can be regarded as basic vectors consisting of retained components for next projection stage[40].

3.3. Least Squares Projection

In this step, a set of features from the original time series dataset X will be obtained. Least squares (LS) is applied to project the original X dataset on to basic vectors $\hat{F}$. But it is not exactly same with classical LS technique. Particularly, the projection procedure is to quantify the correlation among basic vectors, original observation and LS sense value

$$X^T \cdot A = F \Rightarrow A = (X^T)^{-1} F \quad (6)$$

where $X = (x_i, x_2, ..., x_n)$ is a matrix of size $n * m$ which represents the original dataset; $\hat{F}^T = (\hat{f}_1, \hat{f}_2, ..., \hat{f}_k)$ with size $m * k$ corresponds to the standardized basic vectors; $\mathbf{a_i^t} = (a_{i1}, a_{i2}, ..., a_{ik})$ represents the feature vector based on $\mathbf{x_i^t} = (x_{i1}, x_{i2}, ..., x_{im})$.

Finally, orthotropic feature vectors $(a_1^t, a_2^t, ..., a_n^t)$ corresponds to all n time series is obtained. This time-frequency feature vector can represent the original time series dataset to some extent, and in more representative and significant way. Clustering and classification on this dataset result in the grouping together of time series with similarity time-frequency patterns.

4. Experimental Settings

In this section, we give a detailed introduction to experimental settings for the synthetic dataset generation and results evaluation metrics.

4.1. Synthetic dataset generation

The synthetic dataset in this study are based on two basic adding sinusoidal function

$$\lambda_1(t) = sin(\frac{2\pi t}{T}) + \frac{1}{2}sin(\frac{4\pi t}{T} - 2) + \frac{1}{3}sin(\frac{4\pi t}{T} - 2) + \frac{1}{3}cos(\frac{8\pi t}{T} - 2) + sin(\frac{10\pi t}{T} - 6) \quad (7)$$

$$\lambda_2(t) = \frac{2}{3}sin(\frac{2\pi t}{T}) + \frac{1}{2}cos(\frac{12\pi t}{T} - 1) + \frac{1}{6}cos(\frac{4\pi t}{T} - 7) + sin(\frac{4\pi t}{T} - 9) \quad (8)$$

where T is a half of the series length and $t = 1, 2, ..., L$. The objective of mixing many sinusoidal forms is to obtain time series with many peaks. This type of database is relied on Mori's work [10], and the level of noise, outliers is not fixed. The idea is to evaluate the influence for clustering technique on different time series characteristics. In this process, the noise is introduced separately in each series of the database by adding random values issued from a norm distribution of mean 0. The standard deviation (σ) of this distribution is defined by the level of noise normalized by maximum (max) and minimum (min) values of the series:

$$\sigma = \frac{noise_{level}}{max - min} \quad (9)$$

During the dataset generation, the level of noise has four different values $(0, 2, 4, 6)$, and the level 6 represents the highest noise. Meanwhile, the level of outliers is setted in the same way. The format of this synthetic dataset is CSV, and there are 4800 samples in the dataset. Because this dataset is generated by special expressions, the data is grouped into 5 classes. The selected outlier level will represent the proportion of point in the series that will become outliers. Given a specific outlier level, the corresponding number of points are selected randomly from the series and interchanged with points randomly chosen from other series in order to convert them into outliers.

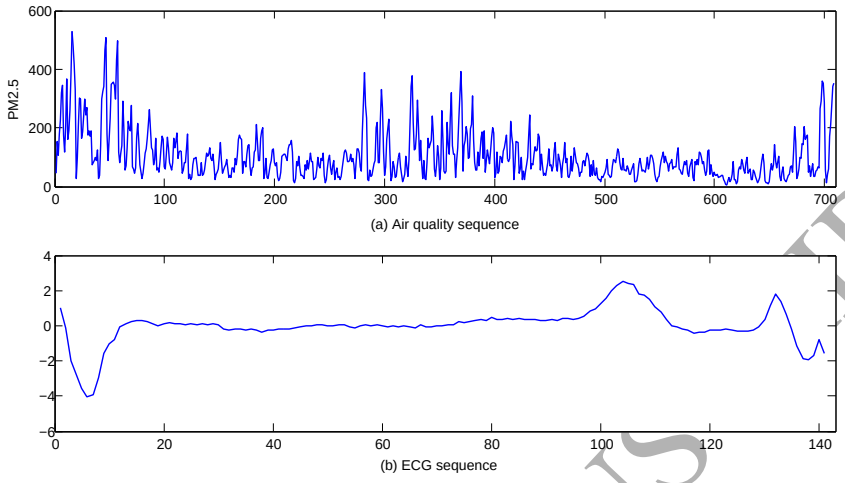

Figure 2: The comparison of Air quality sequence and ECG sequence

4.2. Characteristics of air quality data

Air quality time series contains large amounts of linear and non-linear patterns, which are difficult to analyze [41]. Many studies have been carried out on the basis of air quality data [42, 41]. In this research, the air quality data consists of 31 provincial capitals in China(excluding Hong Kong, Macao, Taiwan). To verify the performance of EPLS, we choose $PM_{2.5}$ and PM_{10} as sample. The database encompasses 709 data sample from January 01, 2014 to December 10, 2015 [43].

We have mentioned that our EPLS has the ability to deal with high noise and outlier aiming at nonlinear and non-stationary time series. From Figure 2 we can see, our $PM_{2.5}$ from Air quality database shows sharp rise, drop and without salient time-frequency patterns[44]. In comparison, the ECG sequence from well-known ECGFiveDays [45] database exhibits strong regularity, seasonality without much mixed frequency characteristic. But the patterns are found out obviously, we do not quantify it based on selected targets.

4.3. Results evaluation metrics

Three metrics are introduced to evaluate the performance of these clustering results, which are the Accuracy metric [46], the $F_{measure}$ and the $RandIndex$. All the metrics are described as follows:

$$Accuarcy = \frac{1}{N}\sum_{i=1}^{N}\hat{L}_i = L_i \tag{10}$$

where N is the total number of instances in the experiment set, L_i the true set of specified labels for instances i and $\hat{L}_i$ is the clustering labels for the same instance.

To calculate the $F_{measure}$ and $RandIndex$, the following quantities should be considered : Let $|TP|$ (True Positive) represents the number of pairs which belongs to one cluster in X (ground truth) and are clustered into Y. The $|TN|$ (True Negative) is the number of pairs, each neither belongs to the same cluster in X, nor clustered into Y. Then $|FN|$ (False Negative) as the error clustering which is the number of pairs belonging to one cluster in X, but not clustered into Y. On contrary, $|FP|$ (False Positive) is the number of pairs which belongs to one cluster in X, but are clustered into Y [47]. The second metric is the $F_{measure}$, which is formulated by,

$$F_{measure} = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{11}$$

$F_{measure}$ is well established to assess the quality of clustering. In which, Precision indicates that how many items in the same cluster are clustered into the same class :

$$Precision = \sqrt{\frac{|TP|}{|TP| + |FP|}} \tag{12}$$

The Recall represents the number of objects in the same class(in ground truth) are clustered into the same cluster.

$$Recall = \sqrt{\frac{|TP|}{|TP| + |FN|}} \tag{13}$$

The third performance metric we have chosen is the RandIndex described as [48]:

$$RandIndex = \sqrt{\frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}} \tag{14}$$

Intuitively, $|TP|+|TN|$ can be considered as the number of agreements between X and Y and $|FP| + |FN|$ as the number of disagreements between X and Y.

5. Experiment Results

The clustering results are presented in this section. To assess the performance of EPLS, the synthetic time series and air quality dataset are applied. We compare the result from four aspects: (1) the adaption for different clustering; (2) time-frequency pattern mining; (3) different distance for database; (4) spatial patterns for air quality data. The results are described in detail.

All experiments are carried out on a desktop computer with following configurations: CPU (Intel Core i7-4770, 3.40GHz); RAM (32GB), Operating System (Windows 7 Professional). The experiments are executed based on Matlab 2014a and RStudio.

5.1. Synthetic database clustering

As Section 4 shows, this synthetic databse consist of 4×4 dataset based on different noise level and outliers level. In every dataset, there are five classes which separated by different shift, p and so on. The dataset is named after $PurityMN$, and M represents the noise level, N represents the outliers level. Results of the synthetic experiments are shown in Table 1. The EPLS and kMeans-R are all relied on Euclidean Distance (ED). EPLS is clustering the results of EPLS feature vectors, otherwise raw time series is clustering by kMeans-R. The performance of the methods presented in this section are measured by the $F_{measure}$ and $RandIndex$. A score of 1 indicates best clustering performance, with 0 corresponding to maximal mixing between the clusters.

The results illustrate that EPLS outperforms raw time series clustering over the 16 datasets. To illustrate why EPLS appears to be performing so well

Table 1: Comparison of clustering results based on synthetic dataset.

	Purity00		Purity02		Purity04		Purity06	
	EPLS	kMeans-R	EPLS	kMean-R	EPLS	kMeans-R	EPLS	kMeans-R
RandIndex	0.94	0.68	1.00	0.67	0.82	0.51	0.95	0.55
F-measure	0.97	0.79	1.00	0.78	0.89	0.57	0.97	0.67
	Purity20		Purity22		Purity24		Purity26	
	EPLS	kMeans-R	EPLS	kMean-R	EPLS	kMeans-R	EPLS	kMeans-R
RandIndex	1.00	0.57	0.96	0.56	0.90	0.71	0.98	0.51
F-measure	1.00	0.69	0.98	0.68	0.95	0.82	0.98	0.59
	Purity40		Purity42		Purity44		Purity46	
	EPLS	kMeans-R	EPLS	kMean-R	EPLS	kMeans-R	EPLS	kMeans-R
RandIndex	0.96	0.69	0.92	0.56	0.92	0.58	0.95	0.55
F-measure	0.98	0.80	0.96	0.64	0.96	0.71	0.97	0.65
	Purity60		Purity62		Purity64		Purity66	
	EPLS	kMeans-R	EPLS	kMean-R	EPLS	kMeans-R	EPLS	kMeans-R
RandIndex	0.92	0.50	0.94	0.58	0.91	0.56	0.91	0.52
F-measure	0.96	0.54	0.97	0.70	0.95	0.68	0.95	0.59

in producing accurate time-frequency features for noisy time series with high level outliers, the effects of varying the outliers level and noise level are also considered. From Table 1 we can see, the results of EPLS is always higher than 0.9 in both RandIndex and $F_{measure}$. Particularly, when the data are noisy with high outliers level, we can also get perfect results based on EPLS. Compared with EPLS, the results of raw time series clustering is not so perfect. Firstly, all the results are not stable, for they are affected seriously by noise and outliers. Secondly, with the increasing of noise level, the performance of kMeans-R show a trend of descending. As the same way, outliers level affects the performance of kMeans-R. These indicate the EPLS has the ability to avoid noise and outliers. Analyzing the theory of EPLS, we get a set of orthogonal basic vectors with several dimensions by applying EEMD algorithm and PCA. This characteristic makes EPLS perform well for clustering.

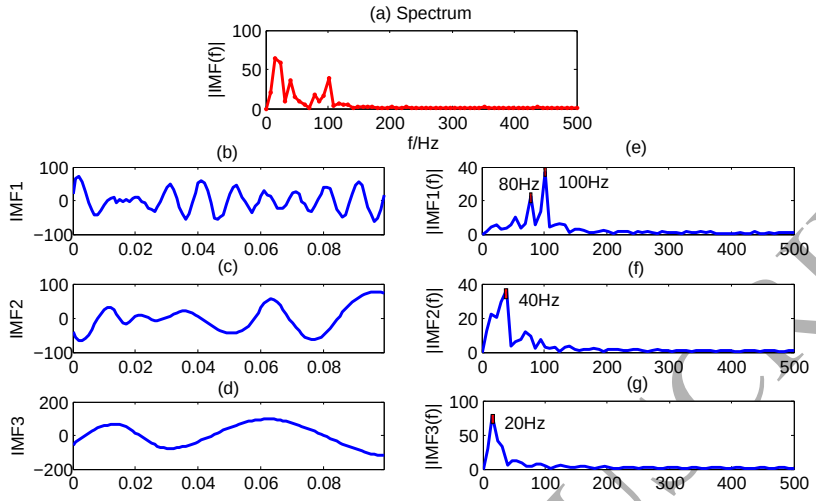

Figure 3: Time-frequency pattern of synthetic time series

5.2. Time-frequency pattern Mining

To some extent, EPLS is effective for clustering and can get stationary performance regardless of the influence of outliers and noise level. In addition to this, EPLS can reveal the time-frequency patterns as Fig.3 shows. The aggregation data is based on $\lambda_1(t)$ which consists of four frequencies, during the generation of dataset, some noise and outliers is added into this group. So this data has high noise and outliers level (both 6 level). Traditional approaches can not well adapt to this dataset, and obtain potential time-frequency patterns. This data has high noise and outliers level (both 6 level). Fig.3(e,f,g) represent the frequency spectrum of time series, and the red dots are the principal frequency corresponding to IMF1, IMF2, IMF3. From Fig.3 (a) we can see, the EEMD technique can reveal the frequency characteristic during the data is noisy with high outliers level. It is notable that EEMD can decompose the time series into some sequences with relative single frequency (see Fig.3 (b,c,d)). From the spectrogram (Fig.3 (e,f,g)) we can see, the first three high frequency IMFs consist of $100Hz$, $80Hz$, $40Hz$, and $20Hz$. So the EPLS algorithm has the ability to get better clustering results and reveal the time-frequency characteristic of

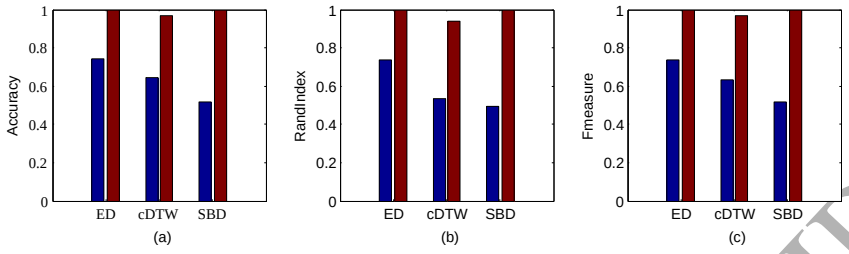

Figure 4: Results based on different distance measures (Red represents the EPLS feature clustering, and blue is the original clustering)

time series.

5.3. The comparison of different distance measures

The high dimensionality, non-stationary, high feature correlation, and large amount of noise, outlier that characterize time series present difficult challenges for clustering. EPLS is proposed to overcome these difficulties. The distance selection of distance measures is probably due to the natural characteristics of database. In a word, time series clustering relies on similarity metrics to a large extent. In this experiment, ED, cDTW and Shape SBD distance measures are taken into account. We evaluate the results from three aspects, *accuracy*, $RandIndex$ and $F_{measure}$.

This experiment is carried out based on kMeans clustering methods combined with ED, cDTW and SBD [6] distance measures. The results are shown in Fig.4. When conducting experiments based on Air quality time series, we can see that EPLS feature clustering is superior to raw time series clustering. For *accuracy*, $RandIndex$ and $F_{measure}$ measures, ED performs best and cDTW follows. But there is no significant difference among these three distance measures based on EPLS, the Fmeasure is above 95%. We can figure out that for raw Air quality time series, ED is the best suitable distance measure. When comparing the results based on EPLS feature, we can see that ED, cDTW and SBD distance measures all achieve high performance. The low dimensional pattern makes EPLS feature clustering distance independent to some extent.

5.4. The adaption of different clustering methods based on Air quality data

To evaluate the performance of EPLS based on Air quality database, Hierarchical clustering [49], Spectral clustering [49] and k-Shape [6] clustering methods are applied. The performance are evaluated by *Accuracy*, $F_{measure}$ and *RandIndex* in this section.

Table 2: Performance comparison of different clustering methods based on Air quality data

	Raw data				EPLS feature vector			
	kMeans	Hierarchical	Spectral	kShape	kMeans	Hierarchical	Spectral	kShape
Accuracy	74.19%	77.42%	93.55%	90.32%	100.00%	98.76%	82.26%	96.77%
RandIndex	73.76%	64.46%	50.03%	49.39%	100.00%	97.65%	55.58%	93.65%
F-measure	73.76%	65.66%	64.41%	62.31%	100.00%	98.76%	71.49%	95.69%

As table 2 shows, our EPLS feature-based clusterings outperforms traditional clustering methods based on Air quality data. For raw data clustering, kMeans performs well with a score above 70%, and the hierarchical follows. But in comparison, the Spectral and kShape clustering methods show a relative unsatisfactory results. The evaluation metrics are all inferior to 70%, and the *RandIndex* metrics are 50% and so for Spectral and kShape methods. When these four clustering methods are applied to EPLS feature vector, we can see a clear boost in performance. Firstly, we can divide the results into four groups by clustering methods.We can see that EPLS-based kMeans obtains a 100% performance over all the three metrics, compared with traditional kMeans the performance increased by 25%. The same as hierarchical and kShape, they all get Accuracy, *RandIndex* and $F_{measure}$ values above 95%. Although the results of EPLS-based Spectral is not so undesirable, it is a little better than traditional Spectral. These traditional clustering methods can achieve high performance in some cases, but for air quality data their performance is not so great. Because traditional clustering methods is not very good at handling dataset with high dimensionality, high level of noise and outlier[2, 10]. Therefore, our EPLS shows good performance for its noise and outlier immunity, and can transfer time series from N-dimension space to relatively low dimensional space for computing conveniently.

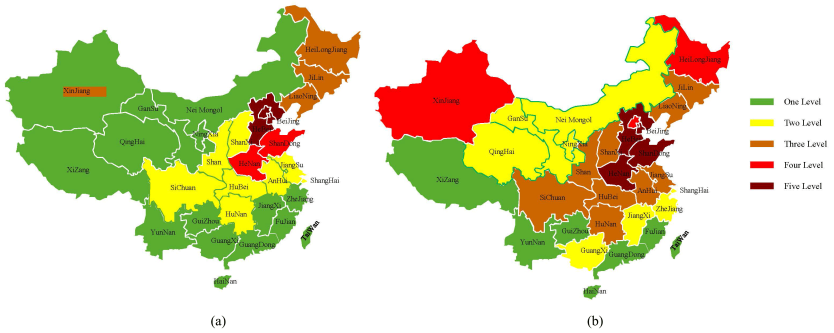

Figure 5: Clustering results based on Air quality databases. (a) EPLS clustering result; (b) AQI clustering result.

5.5. The clustering results of spatial time series

In this section, we explain the spatial feature of EPLS in view of air quality index division. We divide the air quality into five levels (from one level to five level): Good, Moderate, Lightly Polluted, Moderately Polluted and Heavily Polluted. In fact, there are six grades accurately, but in this project we just discuss five grades. So the results have one or two merges.

From Fig.5 we can see, the air quality have different distribution from (a) to (b). Sub-figure (a) cluster air quality databases relying on EPLS, and sub-figure (b) is raw time series clustering. By contrast, some differences can be pointed out. Compared with raw time series, EPLS emerges one level and two level, and the major difference is between XinJiang province and HeiLongJiang province. EPLS is more likely to partition air quality grades considering spatial feature, and it divides geographically closed cities into one grade. This is explainable, usually closed cities have similar air regime. EPLS can discover this regional similarity to a certain extent. From a macro point of view, we can see that Hebei as the center of severe pollution, the air quality becoming better in a radiating outward distribution. This situation is obvious both in Fig.5(a) and Fig.5(b). In a word, our EPLS is meaningful when applied to analyzing air quality. Especially, EPLS can discover this regional similarity to a certain extent and greatly reduce the storage resource.

6. Conclusion and future work

In this paper, we focus on the feature extraction from time series. A novel algorithm named EPLS is proposed for $PM_{2.5}$ time series feature extraction. The EPLS algorithm can obtain a set of feature vectors for data mining techniques. The positive results obtained from experiments demonstrate that when comparing with other clustering methods and distance measures, EPLS is a little superior to cluster time series. Meanwhile, EPLS can decompose non-stationary time series into some sub-series with relative single frequency pattern, even those have high noise and outlier level. By constructing orthogonal basis vectors with PCA, the dimension of original time series is sharply reduced. The above features greatly reveal the potential patterns , and provides convenience for computing. Also, the low dimension of time series reduces the distance measure dependency issues..

Time series has the properties of shift, noise, skewness, dimension, and so on. The first recommended future research work is to compare other clustering methods in various time series. Another proposal for future work includes time series classification. EPLS algorithm are supposed to be used in classification and clustering. Classification based on EPLS is attractive in many fields, and extra advantages of EPLS can be discovered.

7. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China(Nos. 61440018, 61501411), the Hubei Natural Science Foundation(No. 2014CFB904), China Scholarship Council Funding.

References

[1] B. D. Fulcher, N. S. Jones, Highly comparative feature-based time-series classification, IEEE Transactions on Knowledge and Data Engineering 26 (12) (2014) 3026–3037.

[2] D. Barrack, J. Goulding, K. Hopcraft, S. Preston, G. Smith, Amp: a new time-frequency feature extraction method for intermittent time-series data, arXiv preprint arXiv:1507.05455.

[3] S. Askari, N. Montazerin, A high-order multi-variable fuzzy time series forecasting algorithm based on fuzzy clustering, Expert Systems with Applications 42 (4) (2015) 2121–2135.

[4] V. M. A. d. Souza, D. F. Silva, G. E. d. A. P. A. Batista, et al., Extracting texture features for time series classification, in: International Conference on Pattern Recognition, 22nd, International Association of Pattern Recognition-IAPR, 2014.

[5] Y. Sakurai, Y. Matsubara, C. Faloutsos, Mining and forecasting of big time-series data, in: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, ACM, 2015, pp. 919–922.

[6] J. Paparrizos, L. Gravano, k-shape: Efficient and accurate clustering of time series, in: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, ACM, 2015, pp. 1855–1870.

[7] T. Oates, Identifying distinctive subsequences in multivariate time series by clustering, in: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 1999, pp. 322–326.

[8] F. Petitjean, A. Ketterlin, P. Gançarski, A global averaging method for dynamic time warping, with applications to clustering, Pattern Recognition 44 (3) (2011) 678–693.

[9] T. Rakthanmanon, E. J. Keogh, S. Lonardi, S. Evans, Time series epenthesis: Clustering time series streams requires ignoring some data, in: 2011 IEEE 11th International Conference on Data Mining, IEEE, 2011, pp. 547–556.

[10] U. Mori, A. Mendiburu, J. A. Lozano, Similarity measure selection for clustering time series databases, IEEE Transactions on Knowledge and Data Engineering 28 (1) (2016) 181–195.

[11] E. Keogh, S. Kasetty, On the need for time series data mining benchmarks: a survey and empirical demonstration, Data Mining and knowledge discovery 7 (4) (2003) 349–371.

[12] E. Keogh, J. Lin, Clustering of time-series subsequences is meaningless: implications for previous and future research, Knowledge and information systems 8 (2) (2005) 154–177.

[13] L. Ye, E. Keogh, Time series shapelets: a new primitive for data mining, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2009, pp. 947–956.

[14] T. Rakthanmanon, E. Keogh, Fast shapelets: A scalable algorithm for discovering time series shapelets, in: Proceedings of the 13th SIAM international conference on data mining, SIAM, 2013, pp. 668–676.

[15] T. W. Liao, Clustering of time series data a survey, Pattern recognition 38 (11) (2005) 1857–1874.

[16] M. Vlachos, D. Gunopulos, G. Das, Indexing time-series under conditions of noise, Data mining in time series databases 57 (2004) 67–100.

[17] L. Wang, K. Lu, P. Liu, Compressed sensing of a remote sensing image based on the priors of the reference image, IEEE Geoscience and Remote Sensing Letters 12 (4) (2015) 736–740.

[18] Y. Ma, H. Wu, L. Wang, B. Huang, R. Ranjan, A. Zomaya, W. Jie, Remote sensing big data computing: challenges and opportunities, Future Generation Computer Systems 51 (2015) 47–60.

[19] A. Nanopoulos, R. Alcock, Y. Manolopoulos, Feature-based classification of time-series data, International Journal of Computer Research 10 (3) (2001) 49–61.

[20] W. Song, Z. Deng, L. Wang, B. Du, P. Liu, K. Lu, G-ik-svd: parallel ik-svd on gpus for sparse representation of spatial big data, The Journal of Supercomputing (2016) 1–18.

[21] M. Vlachos, S. Y. Philip, V. Castelli, On periodicity detection and structural periodic similarity., in: SDM, Vol. 5, SIAM, 2005, pp. 449–460.

[22] L. Wang, W. Song, P. Liu, Link the remote sensing big data to the image features via wavelet transformation, Cluster Computing 19 (2) (2016) 793–810.

[23] H. Deng, G. Runger, E. Tuv, M. Vladimir, A time series forest for classification and feature extraction, Information Sciences 239 (2013) 142–153.

[24] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, IEEE transactions on knowledge and data engineering 26 (8) (2014) 1819–1837.

[25] Y. Chen, F. Li, Z. Deng, X. Chen, J. He, Pm2. 5 forecasting with hybrid lse model-based approach, Software: Practice and Experience.

[26] M. G. Sefidmazgi, M. M. Kordmahalleh, A. Homaifar, S. Liess, Change detection in climate time series based on bounded-variation clustering, in: Machine Learning and Data Mining Approaches to Climate Science, Springer, 2015, pp. 185–194.

[27] P. Kassomenos, S. Vardoulakis, A. Chaloulakou, A. Paschalidou, G. Grivas, R. Borge, J. Lumbreras, Study of pm 10 and pm 2.5 levels in three european cities: analysis of intra and inter urban variations, Atmospheric Environment 87 (2014) 153–163.

[28] Q. Zhou, H. Jiang, J. Wang, J. Zhou, A hybrid model for pm 2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network, Science of the Total Environment 496 (2014) 264–274.

[29] D. Z. Antanasijevic, V. V. Pocajt, D. S. Povrenovic, M. Ristic, A. A. Peric-Grujic, Pm 10 emission forecasting using artificial neural networks and genetic algorithm input variable optimization, Science of the Total Environment 443 (2013) 511–519.

[30] M. Menzel, R. Ranjan, L. Wang, S. U. Khan, J. Chen, Cloudgenius: a hybrid decision support method for automating the migration of web application clusters to public clouds, IEEE Transactions on Computers 64 (5) (2015) 1336–1348.

[31] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, E. Keogh, Experimental comparison of representation methods and distance measures for time series data, Data Mining and Knowledge Discovery 26 (2) (2013) 275–309.

[32] X. Chen, Y. Chen, A. Y. Zomaya, R. Ranjan, S. Hu, Cevp: Cross entropy based virtual machine placement for energy optimization in clouds, The Journal of Supercomputing (2016) 1–16.

[33] Z. Deng, Y. Hu, M. Zhu, X. Huang, B. Du, A scalable and fast optics for clustering trajectory big data, Cluster Computing 18 (2) (2015) 549–562.

[34] X. Wang, K. Smith, R. Hyndman, Characteristic-based clustering for time series data, Data mining and knowledge Discovery 13 (3) (2006) 335–364.

[35] R. Agrawal, C. Faloutsos, A. Swami, Efficient similarity search in sequence databases, in: International Conference on Foundations of Data Organization and Algorithms, Springer, 1993, pp. 69–84.

[36] D. J. Berndt, J. Clifford, Using dynamic time warping to find patterns in time series., in: KDD workshop, Vol. 10, Seattle, WA, 1994, pp. 359–370.

[37] P. Esling, C. Agon, Time-series data mining, ACM Computing Surveys (CSUR) 45 (1) (2012) 12.

[38] Z. Wu, N. E. Huang, Ensemble empirical mode decomposition: a noise-assisted data analysis method, Advances in adaptive data analysis 1 (01) (2009) 1–41.

[39] J. Shlens, A tutorial on principal component analysis, arXiv preprint arXiv:1404.1100.

[40] L. Wang, R. Ranjan, J. Kolodziej, A. Y. Zomaya, L. Alem, Software tools and techniques for big data computing in healthcare clouds., Future Generation Comp. Syst. 43 (2015) 38–39.

[41] L. A. Díaz-Robles, J. C. Ortega, J. S. Fu, G. D. Reed, J. C. Chow, J. G. Watson, J. A. Moncada-Herrera, A hybrid arima and artificial neural networks model to forecast particulate matter in urban areas: The case of temuco, chile, Atmospheric Environment 42 (35) (2008) 8331–8340.

[42] E. Austin, B. A. Coull, A. Zanobetti, P. Koutrakis, A framework to spatially cluster air pollution monitoring sites in us based on the pm 2.5 composition, Environment international 59 (2013) 244–254.

[43] http://www.aqistudy.cn/.

[44] L. Wang, H. Geng, P. Liu, K. Lu, J. Kolodziej, R. Ranjan, A. Y. Zomaya, Particle swarm optimization based dictionary learning for remote sensing big data, Knowledge-Based Systems 79 (2015) 43–50.

[45] T. U. T. S. C. Homepage., http://www.cs.ucr.edu/ eamonn/time-series-data.

[46] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, H. H. Liu, The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis, in: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, Vol. 454, The Royal Society, 1998, pp. 903–995.

[47] S. Aghabozorgi, T. Y. Wah, Clustering of large time series datasets, Intelligent Data Analysis 18 (5) (2014) 793–817.

[48] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, Machine learning 85 (3) (2011) 333–359.

[49] L. Kaufman, P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis, Vol. 344, John Wiley & Sons, 2009.