J. Parallel Distrib. Comput. 103 (2017) 96–103

Contents lists available at ScienceDirect

J. Parallel Distrib. Comput.

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

EPLS: A novel feature extraction method for migration data clustering

Yunliang Chen [a], Fangyuan Li [a,*], Jia Chen [a], Bo Du [a], Kim-Kwang Raymond Choo [b,c], Houcine Hassan [d]

[a] School of Computer Science, China University of Geosciences, Wuhan, 430070, China
[b] University of Texas at San Antonio, USA
[c] University of South Australia, Australia
[d] Polytechnic University of Valencia, Spain

HIGHLIGHTS

- A numerical feature extraction approach EPLS is proposed.
- EPLS attempts to preserve the most valuable features which are adaptive to different distance measures and different clustering approaches.
- EPLS-based clustering algorithm can scale to large-volumes of data for its dimensionality reduction characteristic.
- EPLS can be efficiently suitable for migration data clustering.

ARTICLE INFO

Article history:
Received 29 June 2016
Received in revised form
23 September 2016
Accepted 17 November 2016
Available online 5 December 2016

Keywords:
Migration data
Feature extraction
EPLS
Clustering
Distance measures

ABSTRACT

Nowadays human activity data such as migration data can be easily accumulated by personal devices thanks for GPS. Analysis on migration data is very useful for society decision. Migration data as non-line time series have the properties of higher noise and outliers. Traditional feature extraction methods cannot address this issue very well because of inherent characteristics. Aiming at this problem, a novel numerical feature extraction approach EPLS is proposed. It is an integration of the Ensemble Empirical Mode (EEMD), Principal Component Analysis (PCA) and Least Square (LS) method. The EPLS model includes (1) Mode Decomposition in which EEMD algorithm is applied to the aggregation dataset; (2) Dimension Reduction 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. Experimental results show that EPLS can overcome the higher noise and outliers based on migration data clustering. Meanwhile, EPLS feature extraction method can achieve high performance compared with several different clustering methods and distance measures.

© 2016 Published by Elsevier Inc.

1. Introduction

Human activities data can be easily collected by personal devices in recent years. An increasing number of human activity data is accumulated for the decision of society [33]. Through human activities data analysis [16], researches can find some interesting patterns. For example, some patterns can be mined from twitter data which are useful to forecast stock price [33]. Air quality index can be used to find PM2.5 patterns [5].

Migration data from Baidu system are used to track human migration activities; for example, during Chinese New Year of 2015, 3.6 billion passenger trips data have been recorded. Baidu map gathers an amount of migration activities dataset from smartphones which own Baidu Maps or other apps using its location-based platform. This dataset is human activity related, which reveals human movement pattern. Analysis on migration is an important aspect of many fields such as economy, traffic, and culture [18].

How to analyze and apply these data is a challenge because the database is time related and has a higher noise and nonlinear level. Usually these data are treated as chaotic time series. A chaotic time series naturally has the properties of high dimension and large data size [23,25]. Usually, clustering is used for exploratory data analysis and act as a major processing step for other tasks [21,24,28]. It can be concluded that clustering is classified into three main branches: (1) whole time series clustering; (2) subsequence clustering; (3) time point clustering [12]. As for whole

* Corresponding author.
E-mail addresses: Cyl_king@hotmail.com (Y. Chen), lfy_cug@hotmail.com (F. Li), jia_2011_cug@aliyun.com (J. Chen), db_cug@yahoo.com (B. Du).

http://dx.doi.org/10.1016/j.jpdc.2016.11.008
0743-7315/© 2016 Published by Elsevier Inc.

time series clustering, there are three different categories, namely shape-based approach [23,46,27], feature-based [2,19] and model-based [44,35]. Feature-based clustering is taken into consideration in this study.

Traditional feature extraction methods usually have some limitation to deal with time series clustering [2,19] because of the nature properties of chaotic time series such as nonlinear, high level of noise and outlier, and non-stationary. So a new representative method is needed to address these problems. It is suggested that not all of these clustering methods and similarity measures are appropriate for every time series databases [41]. Usually Euclidean Distance (ED) can lead to good clustering results as a useful method [11]. But ED measure is not a general method; for some databases, Elastic measure including Dynamic Time Warping (DTW) and Edit Distance can achieve higher performance [6]. Finding a suitable distance measure or a clustering method in specific dataset with best result is difficult. However, a relatively general feature extraction approach for fixed database can deal with this problem to some extent.

In this paper, EPLS as hybrid model-based approach is proposed to extract numerical features for migration data which is gathered from Baidu map engine. EPLS attempts to preserve the most valuable features which are adaptive to different distance measures and different clustering approaches. EPLS approach is based on Ensemble Empirical Mode Decomposition (EEMD) [9], Principal Component Analysis (PCA) [32] and Least Square (LS) method [3] which transfers the original time series into feature space. The proposed method is immune to dataset with higher noise and outliers. Meanwhile, the extracted feature from EPLS has relatively low dimension which shows that it can be adapted to different distance measures and clustering methods.

This paper is organized as follows. In Section 2, the related work is discussed. In Section 3, the details of EPLS algorithm, as well as EEMD, PCA and LS, are shown. Then, some experiment settings and database description are given in Section 4. In Section 5, the EPLS model based approach is applied to Migration data from Baidu system. Section 6 provides a summary of the results and concludes the whole paper.

2. Background and related work

Data sequences which contain explicit information about timing (e.g. $PM_{2.5}$, stock, speech, population migration) can be looked as time series. Large amount of time series appear in almost every discipline [23,26]. With applying clustering methods, some interesting patterns and correlation can be found in the underlying data [8]. Usually time series analysis depends on the choice of techniques and distance measures, in which the target is to find general approach for Migration data from Baidu. Otherwise, handling the noise and outlier is a key problem when clustering time series [43].

2.1. Clustering methods and distance measures

Time series clustering has become an important topic. Motivated by this trend, many clustering approaches and similarity measures continue to develop [31]. In this work, the focus is on the whole time series with a short or meaningful subsequence, not on the long time series [17]. Generally, whole time series clustering includes five classes: Partitioning, Model-based, Grid-based, Hierarchical and Density-based clustering algorithms. Most studies have focused on high dimensional characteristics of time series (by dimensionality reduction), which usually suffer from overlooking of data and inaccurate performance [1].

The choice of distance measure is a key component of time series clustering. Up to now, there are a large number of distance measures proposed, such as Euclidean Distance (ED) [7], Dynamic Time Warping (DTW) [29], Edit Distance for Real Sequence (EDR) [4], Kullback Leibler distance [15] and Longest Common Sub-Sequence (LCSS) [36]. As is well-known that ED and DTW are the most common distance measures in the time series clustering field. ED is known for its efficiency and simpleness [13]. While ED is sensitive to noise and outlier, it cannot deal with time series with different length either. Compared to ED, DTW is a more generalized realization of ED which can solve the shift problem. But DTW requires quadratic computation; therefore, researches should reduce its complexity for computation conveniently [30]. In fact, there is no universal distance measure. The performance is different as time series database changes [40].

2.2. Feature-based clustering

Feature-based representation of time series are used across science. Mean, Standard deviation, Skewness and Kurtosis are used as features by Nanopoulos [20]. These features are statistical values for time-series with low dimensions. Usue et al. introduced a set of features that contains measures of dimension, shift, correlation, seasonality, trend, noise, outliers, autocorrelation, skewness and kurtosis to represent time series [19], and these features form a characteristic vector for clustering. Vlachos et al. used periodic features obtained partly via the direct Fourier decomposition for clustering of MSN query log and electrocardiography time series data [37]. Duncan et al. showed 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 measures (e.g., automutual information, Approximate Entropy, Lempelziv complexity). (2) frequency domain features. These methods are most commonly underpinned by the discrete Fourier or Wavelet transformation of data. (3) time–frequency domain features. These features can be obtained from Hilbert–Huang transformation, Wigner–ville distribution, Cohen time–frequency distribution.

3. Mode decomposition, component analysis and projection (EPLS)

The target of EPLS is to find a base vector for time series database. 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 are serving as inputs for data mining algorithms. EPLS is a way of dealing with features extraction, the outline of which is shown below. Firstly, all time series in a database are pooled into a 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 database are projected to the obtained vectors. There are three stages for implementing the EPLS algorithm including mode decomposition stage, dimension reduction and projection. The algorithm flow is shown in Fig. 1.

3.1. Mode decomposition

Given database includes n time series $X = (x_1, x_2, \ldots, x_n)$, where each time series is a vector of length m $\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{im})$. Then, all the n time series are aggregated as follows. The process is described as shown in Fig. 2.

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

R, as the representation of a database, is the input of EEMD algorithm. EEMD algorithm consists of shifting an ensemble of white noise-added signal and treats the mean as the final true result [45].

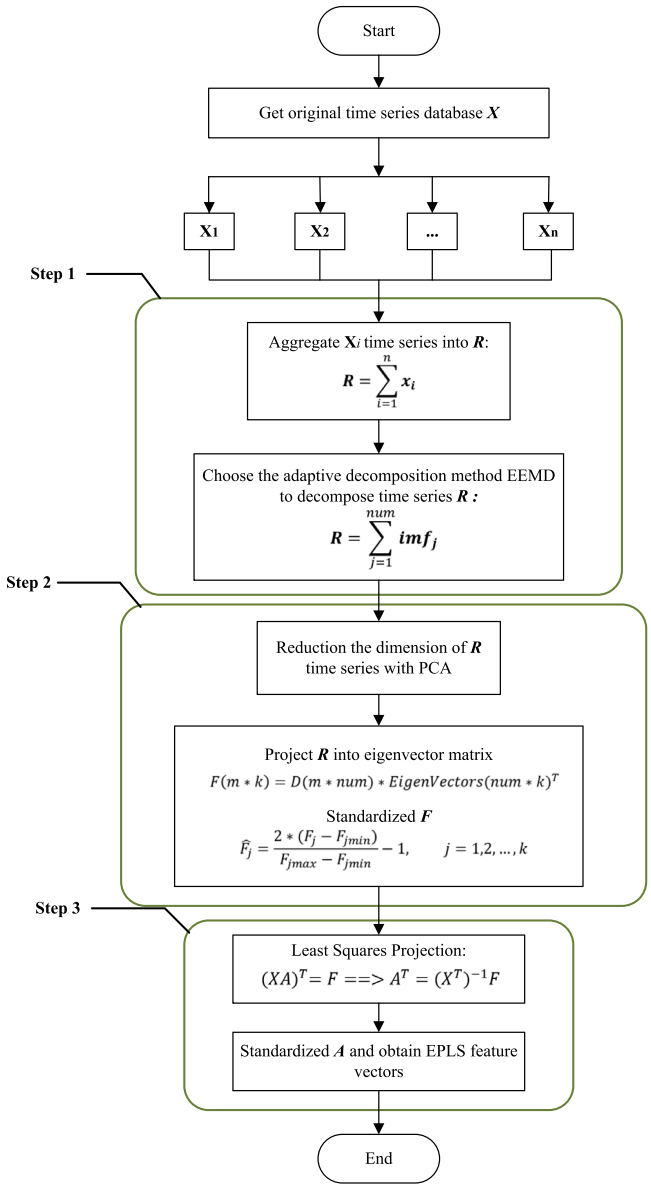


Fig. 1. Flow of EPLS algorithm.

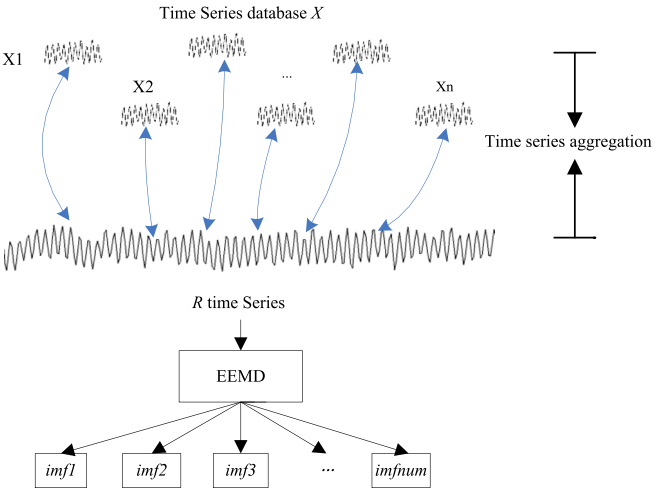


Fig. 2. Mode decomposition process.

Compared with Empirical Mode Decomposition (EMD), EEMD algorithm adds white noise to providing 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) [9]. The EEMD algorithm overcomes two problems: the end effect and the stoppage criteria from EMD algorithm. EEMD signal decomposition technique is used to decompose R into num IMFs

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

subSeries $\mathbf{imf_j} = (imf_{j1}, imf_{j2}, imf_{j3}, \ldots, imf_{jm})$ corresponds to a time–frequency component. Usually, num is less than 10. SubSeries $imf_1, imf_2, \ldots, imf_{num}$ is relevant, and which is ordered 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 features will have a bad effect on next data projecting stage. Principal Component Analysis (PCA) is a standard tool in modern data analysis-in diverse fields from neuroscience to computer graphics [32]. PCA is a simple, non-parametric method for extracting relevant information from confusing databases. In this process, PCA is used for dimension reduction and obtaining totally new orthogonal vectors. The main idea of PCA is to project the $m-dimension$ features into $k-dimension$ features ($K < N$), and the $k-dimension$ features are orthogonal vectors.

Applying PCA technique in subSeries (2) in the following steps: (1) compute the *mean* value of R for every sample, then obtain the new values

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

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

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

So far, the original sample is changed from $num-dimension$ to $k-dimension$ and the dimension reduction is realized. 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. Next, vector $F(m * k)$ is standardized, i.e.

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

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 consisted of retained components for next projection stage.

3.3. Least squares projection

In this step, a set of features from 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 the same with classical LS technique. Particularly, the projection procedure is to

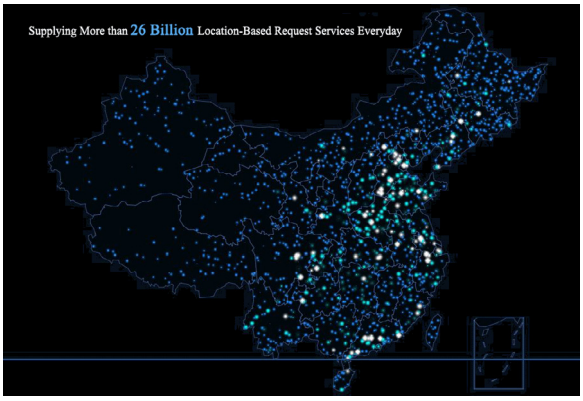


Fig. 3. Location-based data from Baidu [10].

quantify the correlation among basic vectors, original observation and LS sense value

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

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

Finally, orthotropic feature vectors $(a_1^t, a_2^t, \ldots, a_n^t)$ corresponding to all n time series are obtained. This time–frequency feature vector can represent the original time series dataset to some extent. Clustering and classification on this set will result in the grouping together of time series with similarity time–frequency patterns.

4. Experimental settings

Some results of evaluation metrics and the details of experiment database are shown in this section. EPLS is applied to Migration data from Baidu search engine to reveal patterns. The *accuracy*, $F_{measure}$ and *RandIndex* are the evaluation criterion of performance.

4.1. Migration data from Baidu

Population migration time series can reflect the development of economy, technology, health care of cities. In a country, the study of population migration can provide information for urban-planning strategy, massive investment and others.

The movement of people throughout the country during Chinese New Year is recorded by location-based data from search engine Baidu. Baidu launched a heat map revealing where Chinese travelers are coming and heading to, and the most popular migration route which is similar to that shown in Fig. 3.

The Migration data from Baidu used in our study is made up of 31 provincial capitals in China. The migration is from city to city, and every city owes two vectors (the one describes the situation of access to city, called populationIn; the other shows the population from this city to others, called populationOut). As shown, this database has high noise level and outliers level, which reduces the performance in most cases.

4.2. Database characteristic

In order to illustrate that EPLS can deal with database which has high noise and outliers level, noise and outliers need to be defined.

To quantify the noise level of a database, the process is to remove it from every time series in a database [14]. Then, the value of standard deviation need to be computed for each time series. After these steps, we will obtain an overall noise level by statistics metrics. In this work, Discrete Wavelet transform denoising method is applied [38]. In this case, Daubechies wavelet is suitable for time series noise reduction. We have implemented this method with *wavedec* function and *ddencmp* function in Matlab.

Outlier deviated from database has a great effect on some distance measures [22]. Boxplot method is used to calculate the percentage of points which lies outside the whiskers [39].

4.3. Results evaluation metrics

To evaluate the performance of all these clustering results, we will use three metrics as follows. The first is the Accuracy metric [47], which measures the proportion of correct clustering instances:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\hat{L}_i = L_i \tag{7}$$

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

The second metric is the $F_{measure}$ calculated as

$$F_{measure} = \frac{2 \cdot precision \cdot recall}{precision + recall}. \tag{8}$$

The third performance metric we have chosen is the RandIndex described as follows [14]: given a set of n elements $S = (o_1, o_2, \ldots, o_n)$ and two partitions of s to compare, $X = (x_1, x_2, \ldots, x_r)$, a partition of S into r subsets, and $Y = (Y_1, Y_2, \ldots, Y_s)$, a partition of S into s subsets, define the following: a, the number of pairs of elements in S that are in the same set in X and in the same set in Y; b, the number of pairs of elements in S that are in different sets in X and in different sets in Y; c, the number of pairs of elements in S that are in the same set in X and in different sets in Y; d, the number of pairs of elements in S that are in different sets in X and in the same set in Y; The RandIndex can be defined as

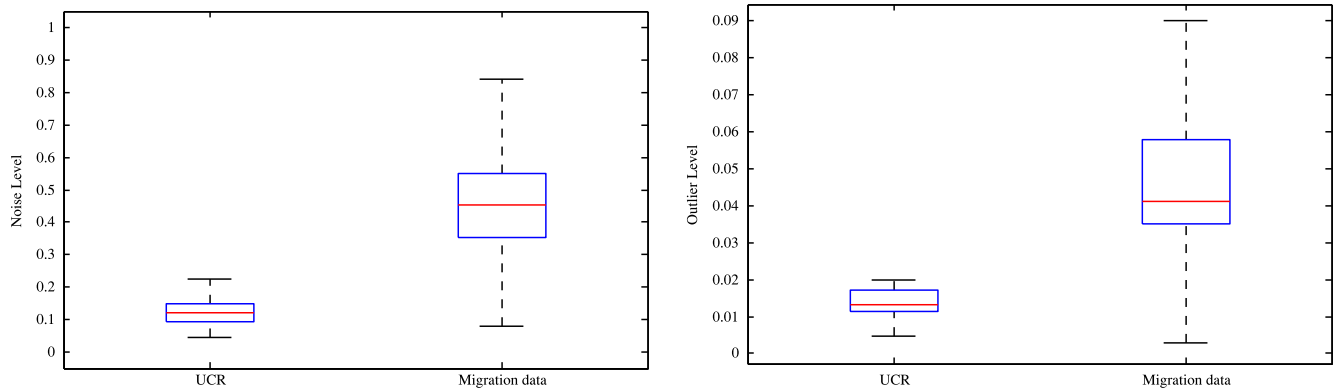
$$RandIndex = \frac{a+b}{a+b+c+d}. \tag{9}$$

Intuitively, $a + b$ can be considered as the number of agreements between X and Y, and $c + d$ as the number of disagreements between X and Y.

5. Experiment results

The clustering results are presented in this section. Migration data from Baidu is applied to EPLS model-based approach. We compare the result from four aspects: (1) the quantization of noise and outlier, (2) the adaption for different clustering methods, (3) different distance measures for database and (4) time–frequency pattern. The results show that EPLS is more effective than traditional frequency and time domain based clustering method. Otherwise, EPLS performs well when the database is with high noise level and outlier level.

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



(a) Noise level. (b) Outlier level.

Fig. 4. The quantization of noise and outlier of databases.

5.1. Characteristics of databases

We have mentioned that our EPLS has ability to deal with high noise and outlier aiming at non-linear and non-stationary time series.

We describe the difference between Migration data from Baidu and UCR databases [34] in Fig. 4. For noise level, wavelet transform is used to de-noising and we quantify the noise by statistic values. We calculate the outlier level based on boxplot technique (details in Section 4.2). In Fig. 4(a), presents the noise level, we can obviously find that Migration data from Baidu has relatively high noise level compared with UCR databases; also, Migration data from Baidu has more higher level of outlier. These results verify that Migration data from Baidu is more complex. Usually clustering is critically depending on the choice of distance measure, and the key issue of which is to handle the variety of distortions, noise and outliers.

5.2. Different clustering methods for Migration data

Clustering is one of the most popular data mining methods. For evaluating the performance of EPLS, there are four clustering methods applied in this section [42]. KMeans and Hierarchical clustering as traditional techniques are taken into consideration, and they are all based on ED measure. For hierarchical clustering, the single, average, complete are the most widely used linkage variants [10], and we adopt complete linkage in experiment. Spectral clustering has attracted more attention due to its significant performance over other types of data [10]. KShape is a novel algorithm proposed by John [23] which is a shape-based time series clustering. The performance are evaluated by *Accuracy*, $F_{measure}$ and *RandIndex* in this section. A score of 1 indicates best clustering performance, with 0 corresponding to maximal mixing among the clusters.

According to the analysis on Migration data, Population in time series and Population out time series have very similar time–frequency characteristics. This makes it difficult to classify these two classes. Therefore, we label the Migration data as PopulationIn and PopulationOut. As Table 1 shows, our EPLS feature-based clusterings are outperforming over traditional clustering methods based on Migration data from Baidu. For EPLS-based Spectral and kShape clustering method, our algorithm obtains 100% *accuracy*. Compared with our EPLS, the raw data based clustering just has *accuracy* and *RandIndex* with 50%, and $F_{measure}$ with 60%. This illustrates that kMeans, Hierarchical clustering, Spectral and kShape are not suitable for Migration data from Baidu Engine. The reason may be that these clustering methods is inappropriate to the dataset with higher noise and outliers. On the other hand, EPSL numerical feature extraction model-based approach is relative universal for these four clustering methods, and it can fit well into Migration data from Baidu.

In this section, we have conducted a series of experiments to evaluate the effectiveness of EPLS for Migration data from Baidu. From above we can see, our EPLS-based clustering algorithm can scale to large volumes of data for its dimensionality reduction characteristic. As to the characteristics of our databases, the nature of non-linear and non-stationary is taken into consideration. EPLS is not limited to noise, outlier, non-stationary, which is mainly due to the properties of EEMD decomposition.

5.3. Different distance measures for Migration data from Baidu

EPLS is proposed to overcome high dimensionality, non-stationary, high feature correlation and large amount of noise, outlier for human related time series. As Section 5.2 shows, EPLS is effective for clustering and can get stationary performance regardless of the influence of outlier and noise level. What is more, EPLS is clustering methods irrelevant to some extent.

For time series clustering, we should define a distance measure which can estimate the similarity among time series. However, there is no universal distance measure. For example, ED measure and the other common measures are not suitable to evaluate the non-temporal time series with noise and outlier. This is probably due to the natural characteristics of databases, which makes only a few distance measures suitable. In a word, time series clustering relies on similarity metrics to a great extent. In this experiment, ED, cDTW and SBD distance measures are taken into account for Migration data from Baidu. 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 [23] distance measures. The results are shown in Fig. 5. We can see that EPLS results are all above 95% among these three distance measures, and the difference is not so obvious. On the contrary, the raw time series clustering do not perform very well based on our Migration database. For *Accuracy*, *RandIndex* and $F_{measure}$ metrics, EPLS results are close to 1 for our Migration data from Baidu. So we can say that EPLS makes Migration data from Baidu adaptive for these three distance measures to some extent.

5.4. The clustering results and time–frequency pattern

In this section, we show the clustering result of EPLS numerical features from Migration data. This EPLS feature has three dimensions. From the three dimensional representation (Fig. 6) of the feature values, the individuals from two groups are separated

Table 1
Performance comparison of different clustering methods based on Migration data.

	Raw data				EPLS feature vector			
	kMeans	Hierarchical	Spectral	kShape	kMeans	Hierarchical	Spectral	kShape
Accuracy	53.23%	51.61%	57.45%	58.06%	98.39%	96.77%	100.00%	100.00%
RandIndex	49.39%	49.23%	50.23%	51.30%	96.77%	97.65%	100.00%	100.00%
F-measure	61.73%	66.30%	64.31%	57.67%	98.39%	93.65%	100.00%	100.00%

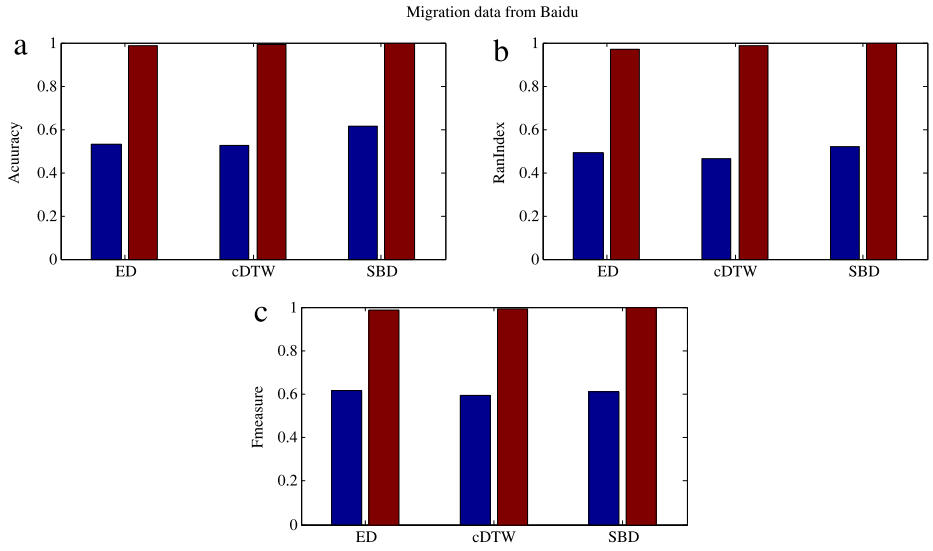


Fig. 5. Results based on different distance measures. (a) *Accuracy* measure results; (b) *RandIndex* measure results; (c) $F_{measure}$ measure results; the red line represents our EPLS results, the blue is raw time series results. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

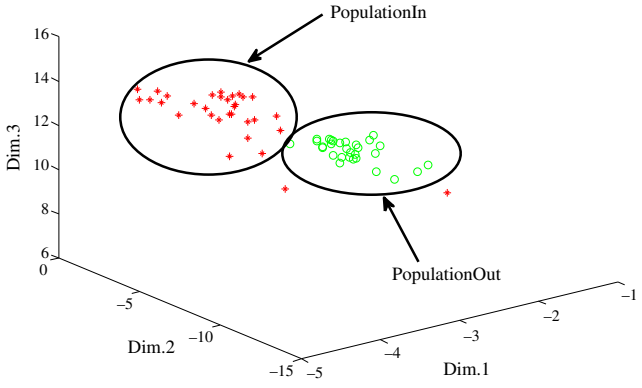


Fig. 6. kMeans clustering result for Migration data from Baidu.

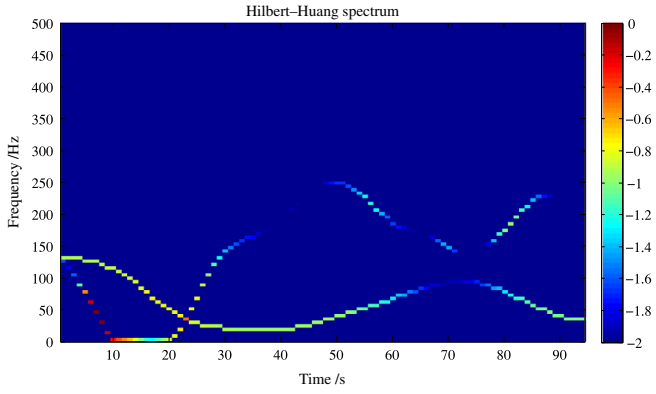


Fig. 7. EEMD time–frequency of Migration data from Baidu.

from each other. In this experiment, the dimension of Migration data reduces from 74 to 3, but reserves enough information for clustering. This is meaningful for further studying.

Fig. 7 is the Hilbert–Huang spectrum based on EEMD algorithm, which clearly explains that Migration data from Baidu have two time–frequency patterns. This also shows that EPLS can reveal inner characteristics from database, and then provide support for precise numerical feature extraction. In a word, EEMD is efficient for time series time–frequency analysis; PCA offers a series of features which is linearly irrelevant; LS projects raw time series into feature space. These three procedures are all linked with each another, and these procedures cannot be out-of-order.

6. Conclusion and future work

The study of feature extraction for Migration data is significant for urban planning and population research. In this paper, a novel numerical feature extraction method EPLS is proposed to address this problem. A series of experiments have been carried out to verify that proposed EPLS is valuable. It also has the society meaning; for example, the analysis on Baidu Migration data can provide useful suggests for government decision in Chinese New year Season.

Firstly, the noise level and outlier level are quantified to ensure that EPLS can be immune from noise and outlier. Secondly, a set of experiments are performed to illustrate that EPLS can adapt to different clustering situations. It can also be revealed that extracted EPLS feature is suitable for ED, cDTW and SBD distance measures. Finally, the clustering results discover the reasons why EPLS can achieve better results.

In future, we have two different directions for further study. Primarily, we will perform various time series to mining their diverse properties such as shift, entropy, and skewness. The other proposal is to classify time series. Additionally, Migration data from Baidu has significant scientific meaning; we can study the

correlation and effect with other attributes such as economy and environment.

References

[1] S. Aghabozorgi, T.Y. Wah, Clustering of large time series datasets, Intell. Data Anal. 18 (5) (2014) 793–817.
[2] D. Barrack, J. Goulding, K. Hopcraft, et al. AMP: a new time-frequency feature extraction method for intermittent time-series data, 2015. arXiv preprint arXiv:1507.05455.
[3] A. Charnes, E.L. Frome, P.L. Yu, The equivalence of generalized least squares and maximum likelihood estimates in the exponential family, J. Amer. Statist. Assoc. 71 (353) (1976) 169–171.
[4] L. Chen, M.T. ?zsu, V. Oria, Robust and fast similarity search for moving object trajectories, in: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, ACM, 2005, pp. 491–502.
[5] W.G. Cobourn, An enhanced PM 2.5 air quality forecast model based on nonlinear regression and back-trajectory concentrations, Atmos. Environ. 44 (25) (2010) 3015–3023.
[6] H. Deng, G. Runger, E. Tuv, M. Vladimir, A time series forest for classification and feature extraction, Inform. Sci. 239 (2013) 142–153.
[7] C. Faloutsos, M. Ranganathan, Y. Manolopoulos, Fastsubsequence matching intime-seriesdatabases, ACMSIGMOD Rec. 23 (1994) 419–429.
[8] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, J. Intell. Inf. Syst. 17 (2) (2001) 107–145.
[9] N.E. Huang, Z. Shen, S.R. Long, et al., The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 454 (1971) (1998) 903–995. The Royal Society.
[10] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Vol. 344, John Wiley & Sons, 2009.
[11] E. Keogh, S. Kasetty, On the need for time series data mining benchmarks: a survey and empirical demonstration, Data Min. Knowl. Discov. 7 (4) (2003) 349–371.
[12] E. Keogh, J. Lin, Clustering of time-series subsequences is meaningless: implications for previous and future research, Knowl. Inf. Syst. 8 (2) (2005) 154–177.
[13] E. Keogh, M. Pazzani, K. Chakrabarti, S. Mehrotra, A simple dimensionality reduction technique for fast similarity search in large time series databases, Knowl. Inf. Syst. 1805 (2000) 122–133.
[14] T. Köhler, D. Lorenz, A Comparison of Denoising Methods for One Dimensional Time Series, University of Bremen, Bremen, Germany, 2005, p. 131.
[15] M. Kumar, N.R. Patel, J. Woo, Clustering seasonality patterns in the presence of errors, in: Proceedings of KDD 02, Edmonton, Alberta, Canada.
[16] Y. Li, W. Dai, Z. Ming, et al., Privacy protection for preventing data over-collection in smart city, IEEE Trans. Comput. 65 (5) (2016) 1339–1350.
[17] J. Lin, D. Etter, D. DeBarr, Exact and approximate reverse nearest neighbor search for multimedia data, in: International Conference on Data Mining, 2008, pp. 656–667.
[18] B. Minor, J.R. Doppa, D.J. Cook, Toward Learning and Mining from Uncertain Time-Series Data for Activity Prediction.
[19] U. Mori, A. Mendiburu, J. Lozano, Similarity Measure Selection for Clustering Time Series Databases.
[20] A. Nanopoulos, R. Alcock, Y. Manolopoulos, Feature-based classification of time-series data, in: Information Processing and Technology, Nova, Commack, NY, USA, 2001, pp. 49–61.
[21] 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.
[22] S.H. Pal, J.N. Patet, Time-series data mining: A review, Binary J. Data Min. Netw. 5 (1) (2015) 01–04.
[23] J. Paparrizos, L. Gravano, k-Shape: Effcient and accurate clustering of time series, in: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, ACM, 2015, pp. 1855–1870.
[24] F. Petitjean, A. Ketterlin, P. Gancarski, A global averaging method for dynamic time warping, with applications to clustering, Pattern Recognit. 44 (3) (2011) 678–693.
[25] M. Qiu, Z. Chen, Z. Ming, et al. Energy-aware data allocation with hybrid memory for mobile cloud systems. 2014.
[26] M. Qiu, Z. Chen, J. Niu, et al., Data allocation for hybrid memory with genetic algorithm, IEEE Trans. Emerging Top. Comput. 3 (4) (2015) 544–555.
[27] T. Rakthanmanon, E. Keogh, Fast shapelets: A scalable algorithm for discovering time series shapelets, in: Proceedings of the Thirteenth SIAM Conference on Data Mining, SDM, 2013.
[28] T. Rakthanmanon, E.J. Keogh, S. Lonardi, et al., Time series epenthesis: clustering time series streams requires ignoring some data, in: 2011 IEEE 11th International Conference on Data Mining, (ICDM), IEEE, 2011, pp. 547–556.
[29] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Trans. Acoust. Speech Signal Process. 26 (1978) 43–49.
[30] S. Salvador, P. Chan, Toward accurate dynamic time warping in linear time and space, Intell. Data Anal. 11 (2007) 561–580.
[31] D.E. Seborg, S. Barbara, Clustering of multivariate time-series data, in: Proceedings of the American Control Confersence, vol. 5, 2002, pp. 3931–3936.
[32] J. Shlens, A tutorial on principal component analysis, 2014. arXiv preprint arXiv:1404.1100.
[33] M. Skuza, A. Romanowski, Sentiment analysis of Twitter data within big data distributed environment for stock prediction, in: 2015 Federated Conference on Computer Science and Information Systems, FedCSIS, IEEE, 2015, pp. 1349–1354.
[34] The UCR Time Series Classification/Clustering Homepage. http://www.cs.ucr.edu/~eamonn/time-series-data.
[35] M. Vlachos, D. Gunopulos, G. Das, Indexing time-series under conditions of noise, in: M. Last, A. Kandel, H. Bunke (Eds.), DataMing in Time Series Databases, World Scientific, Singapore, 2004, p. 67.
[36] M. Vlachos, G. Kollios, D. Gunopulos, Discovering similar multidimensional trajectories, in: 2002 Proceedings 18th International Conference on Data Engineering, IEEE, 2002, pp. 673–684.
[37] M. Vlachos, S.Y. Philip, V. Castelli, On periodicity detection and structural periodic similarity, in: SIAM International Conference on Data Mining, vol. 5, SIAM, 2005, pp. 449–460.
[38] Lizhe Wang, Hao Geng, Peng Liu, Ke Lu, Joanna Kolodziej, Rajiv Ranjan, Albert Y. Zomaya, Particle Swarm Optimization based dictionary learning for remote sensing big data, Knowl.-Based Syst. 79 (2015) 43–50.
[39] Lizhe Wang, Shiyan Hu, Gilles Betis, Rajiv Ranjan, A computing perspective on smart city [Guest Editorial], IEEE Trans. Comput. 65 (5) (2016) 1337–1338.
[40] Lizhe Wang, Ke Lu, Peng Liu, Compressed sensing of a remote sensing image based on the priors of the reference image, IEEE Geosci. Remote Sens. Lett. 12 (4) (2015) 736–740.
[41] 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 Min. Knowl. Discov. 26 (2) (2012) 275–309.
[42] Lizhe Wang, Rajiv Ranjan, Joanna Kolodziej, Albert Y. Zomaya, Leila Alem, Software tools and techniques for big data computing in healthcare clouds, Future Gener. Comput. Syst. 43–44 (2015) 38–39.
[43] Lizhe Wang, Weijing Song, Peng Liu, Link the remote sensing big data to the image features via wavelet transformation, Cluster Comput. 19 (2) (2016) 793–810.
[44] T. Warrenliao, Clustering of time series dataa survey, Pattern Recognit. 38 (11) (2005) 1857–1874.
[45] Z. Wu, N.E. Huang, Ensemble empirical mode decomposition: a noise-assisted data anal- ysis method, Adv. Adapt. Data Anal. 1 (01) (2009) 1–41.
[46] 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.
[47] M.L. Zhang, Z.H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1819–1837.

Further reading

[1] https://en.wikipedia.org/wiki/Rand-index.
[2] http://qianxi.baidu.com/.

Yunliang Chen received the B.Sc. and M.Eng degrees from China University of Geosciences, and the Ph.D.degree from Huazhong University of Science and Technology, China. He is currently an associate professor with School of Computer Science, China University of Geoscience, Wuhan, China.

Fangyuan Li received the B.Sc. degree from China University of Geosciences. Currently, she is a graduate student with the School of Computer Science, China University of Geosciences, Wuhan, China.

Jia Chen received the B.Sc. degree from China University of Geosciences. He is currently a postgraduate with School of Computer Science, China University of Geosciences, Wuhan, China. His research interests include data mining and high performance computing.

Bo Du received the B.Sc. degree from China University of Geosciences. He is currently a postgraduate with School of Computer Science, China University of Geosciences, Wuhan, China. His research interests include data mining and high performance computing.

Kim-Kwang Raymond Choo is with Department of Information Systems and Cyber Security, University of Texas at San Antonio, USA, and School of Information Technology and Mathematical Sciences, University of South Australia, Australia.

Houcine Hassan is with Department of Systems Data Processing and Computers Organization, Polytechnic University of Valencia, Spain.