Time Series Clustering Method Based on Principal Component Analysis

Danyang CAO[1, a], Yuan TIAN[1, b], Donghui BAI[2]

[1]College of Computer, North China University of Technology, Beijing 100144 China

[2]CASKY eTech Co.,Ltd

[a]email: ufocdy@163.com, [b]email:448588273@qq.com

Keywords: time series dataset; clustering method; principal component analysis (PCA); Euclidean distance

Abstract. In terms of existing time series clustering method based on Euclidean distance metric, with the increasing dimension of time series, the time complexity of the algorithm will be increased too; and this method can also lead to incorrect clustering result because of it unable to recognize the abnormal values in time series. Principal component analysis retains large variance and contains more information by linear transformation; it can effectively reduce the dimension of the time series and identify outliers. This paper proposes the idea of time series clustering analysis method based on principal component analysis. Firstly, applying principal component analysis to time series dataset, by way of dimension reduction, obtained the corresponding coefficient matrix and eigenvalues. Secondly, using clustering method based on Euclidean distance on the calculated coefficient matrix, the clustering result of coefficient matrix is consistent with time series dataset. Using simulation data and meteorological data to validate this method, the experimental results show that time complexity of time series clustering method proposed in this paper is significantly better than the algorithm based on Euclidean distance, especially for time series dataset which has linear correlation.

1. Introduction

Time series is an important high-dimensional data type, a random sequence which is in accordance with order time and exists interrelationship, widely in the field of engineering, scientific observation, and economic management. Such as the medical heart beat series; weather daily precipitation sequence; the stock closing price series, observation sequence of blackbody on satellites etc [1]. In recent years, more and more scholars began to pay attention to the time series data mining. Time series data mining has very important practical value, time series clustering is an important task among time series data mining.

Different from traditional clustering analysis, the time series data which processed by cluster analysis is time-varying; the aim of time series clustering is to put the similar time series together, and take different time series apart. However, due to the main characteristics of time series: high-dimension, complex, dynamic and high noise etc, making many traditional clustering algorithms cannot be applied to the time series data. At present, time series clustering methods can be divided into three major categories, 1) Raw-data-based approaches, clustering the original time series data directly without any pretreatment, the method mostly used is to change the existing static data clustering method for processing time series. Euclidean distance is the simplest and most widely used method for the numerical time series. Furthermore Manhattan's p=1 and Maximum’s Lp paradigm of $p=\infty$ [2] is also commonly used clustering methods based on the raw time series data. In order to solve the problem of varying lengths of template matching, Berndt etc proposes a dynamic Time warping distance (DTW)[3], but the amount of calculation time is too large and the method commonly used in voice recognition. 2) Feature-based approaches, Due to characteristics of the time series, directly using time series data mining not only cost too much in computing and storage but also affect the reliability and accuracy of data mining algorithms, may be result in "disaster of dimensionality." To avoid these problems, it is necessary to find an appropriate data representation to simplify the dimension of time series data. Commonly used methods are PAA which is proposed by Keogh and Yi [4], Pratt and Fink proposed a piecewise linear representation based on important points [5]. Eammon, who gives a new approach named symbolic aggregate approximation (SAX) on the basis of PAA [6, 7]. SAX defines a distance measure between the symbols, and the method to ensure that the distance between the two symbol sequences to meet the actual distance between the two time series of the lower bound of the requirements, implementation of the method is simple and can effectively solve the problem of high dimensionality. But the downside is that it cannot effectively distinguish the similarity between sequences. 3) Model-based approaches, the basic idea of this category is time series modeling first, and then by using the model parameters, coefficients, and statistical information such as fitting residual to carry out time series clustering. Commonly used time series model is autoregressive moving average model (ARMA), hidden Markov model (HMM), and artificial neural network model (ANN) etc. There are many research results in this aspect, Xiong and Yeung [8] assume that the time series is generated by the K kept ARMA model, use an algorithm called expectation-maximization to learn model coefficients and parameters, compare to other algorithm, this EM algorithm can adaptively determine the optimal clustering number. Oates etc [9] presented a hybrid time series clustering algorithm that uses dynamic time warping and hidden Markov model induction. The two methods complement each other: DTW produces a rough initial clustering and the HMM removes from these clusters the sequences that do not belong to them. The downside is that the HMM removes some good sequences along with the bad ones.

At present the existing methods have their own advantages and disadvantages, Raw-data-based approaches are relatively simple and can be carried out directly on the time series clustering, but this kind of method will not applicable when the size of time series data is too large and with some noise. Feature-based time series clustering approaches by means of reduce dimension to enhance efficiency of the algorithm, but sometimes it will lost important information of the original data after reduction. The advantages of model-based clustering approaches is that can handle time series which has noisy, uncertain characteristic, or too much change, but its disadvantage is that the process is too complicated, requiring establish a model for time series first, then using the model parameters, coefficients etc to excavate the time series. On this basis, this paper proposes a time series clustering method based on principal component analysis, this method combines the advantages of the raw-based approaches and feature-based approaches. The step of this method is applying principal component analysis to time series dataset firstly, by way of dimension reduction, obtained the corresponding coefficient matrix and eigenvalues. Secondly, using clustering method based on Euclidean distance on the calculated coefficient matrix, finally we find that the clustering result of coefficient matrix is consistent with time series dataset. This method can not only obtain accurate clustering result, and compared with the time series clustering method based on Euclidean distance, can significantly improve the efficiency of the algorithm.

This paper is organized as follows. In section 2 we will review the related definitions which used in this paper. In section 3 we will introduce our clustering method and simulation, followed by the experimental evaluations in section 4. The conclusions are given in section 5.

2. Related definitions

2.1. Time Series dataset

Time series is an ordered set of elements combined with record value and record time, recorded as $x = \left\{x(1) = \left(v_1, t_1\right), x(2) = \left(v_2, t_2\right), L, x(T) = \left(v_T, t_T\right)\right\}$, element $\left(v_T, t_T\right)$ indicates the record value v_T at record time t_T [10]. When m time series which has correlation compose a matrix called time series dataset, recorded as $X = \left\{x_1(t), x_2(t), \cdots, x_m(t)\right\}, t = 1, 2, \cdots, T$, it also can expressed as formula (1):

$$X = \begin{pmatrix} x_1(1) & x_1(2) & \cdots & x_1(T) \\ x_2(1) & x_2(2) & \cdots & x_2(T) \\ \vdots & \vdots & \cdots & \vdots \\ x_m(1) & x_m(2) & \cdots & x_m(T) \end{pmatrix} \quad (1)$$

For generalized time series dataset, record values $x_I(t), t = 1, 2, \cdots T$ can be various types, including discrete symbol, structure data, multimedia data, etc. This paper considers only narrow time series, which means $x_I(t), t = 1, 2, \cdots T$ is a real value.

2.2. Principal Component Analysis

Principal component analysis (PCA) [11] is an effective method of multivariate statistical data analysis, on the premise of little or no loss of the raw information, it can converse complex data which has correlation to less number and unrelated integrated indicators to each other by linear combination, and thus it will achieve the goal of dimension reduction. With this idea, formula (1) can approximately express as n of T-dimensional linear independent time series dataset, namely:

$$\{x_1(t), x_2(t), \cdots, x_m(t)\}^T = A\{y_1(t), y_2(t), \cdots, y_n(t)\}^T, t = 1, 2, \cdots, T \quad (2)$$

In this expression, where $A = (a_{ij})_{m \times n}, 1 \le n \le m, n \le T$ is called the coefficient matrix of the raw time series dataset.

2.3. Euclidean distance

Euclidean distance, it is a commonly used distance definition. The Euclidean distance equation between two of n-dimensional vector is:

$$Dis = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \quad (3)$$

Or $$Dis = \sqrt{(a-b)(a-b)^T} \quad (4)$$

3. Time series clustering method based on principal component analysis

3.1. Analysis of Algorithm

Among the current existing time series clustering algorithm, method based on Euclidean distance is simple and widely used, especially feasible for numerical time series. But when the dimension of time series dataset is too large will cause the low efficiency of the algorithm as well as the problem of inaccurate result. Due to the time series dataset which is need to do clustering analysis is made up by several components of different ways, these ingredients contain different information lead to some become to the primary components and another become to secondary components. Combined with the ideology of principal component analysis in multivariate statistical analysis, can identify the main factors of the components and remove some minor factors. Experiments found that a linear combination of time series dataset under these major factors namely coefficient matrix, its Euclidean distance clustering result reflects the clustering result of raw time series dataset.

3.2. Time Series' Clustering Algorithm based on principal component analysis

Input of the algorithm: $\{x_1(t), x_2(t), \cdots, x_m(t)\}, t = 1, 2, \cdots, T$, time series dataset; threshold, Euclidean distance threshold;

Output of the algorithm:

$result = \{R1 \cup R2 \cup \cdots Rn\}, Rn = \{x_i(t) \cup x_j(t) \cdots x_k(t)\}(i, j, k = 1, 2, \cdots m; t = 1, 2, \cdots T)$, time series clustering result;

Specific steps:

Step 1: time series standardization;

```
    $m = mean(X_m(t), axis = 0)$
    $data- = m$
Step2: solving the covariance matrix;
    $C_{m\times t}$ = cov(transpose(data))
Step3: computing eigenvalues and eigenvectors (coefficient matrix);
    $\text{eigvalue, eigvectors}_{n\times t} = \text{linalg.eig(C)}$
Step4: calculate Euclidean distance of coefficient matrix;
    $\text{dis = pdist(eigvectors, 'euclidean')}$
Step5: get the Euclidean distance matrix based on Euclidean distance formula;
    $\text{dis} = \{d_{ij}\}, i = 1, 2 \cdots n, j = i + 1$
Step6: in accordance with the input threshold and calculated matrix in step5, obtain the final result;
    $result = \{R1 \cup R2 \cup \cdots Rn\}, Rn = \{x_i(t) \cup x_j(t) \cdots x_k(t)\}(i, j, k = 1, 2, \cdots m; t = 1, 2, \cdots T)$
```

3.3. Simulation data validation

To validate the accuracy and validity of the proposed method on the clustering result, we design a group of simulation data; it is a set of time series, the expression shown in Table 1: where rand is an integer from 0 to 1.

Table1: a set of simulation time series

Time series	A new time series
$y_1(t) = sign(\cos(t))$	$y_5(t) = y_1(t) + 3 + 2 \times rand$
$y_2(t) = \text{mod}(2\pi t, 1)$	$y_6(t) = y_2(t) + 6 + 0.5 \times rand$
$y_3(t) = \sin(2\pi t / 20)$	$y_7(t) = y_3(t) + 4 + 0.7 \times rand$
$y_4(t) = 4 \times y_1(t) + y_2(t) + 2 \times y_3(t)$	$y_8(t) = y_4(t) + 2 + 3 \times rand$

Above structure of simulation data in table1, can be clearly classify the time series to 4 sets. It is $\{y_1(t), y_5(t)\} \cup \{y_2(t), y_6(t)\} \cup \{y_3(t), y_7(t)\} \cup \{y_4(t), y_8(t)\}, t = 1, 2, 3 \cdots, 100$, the image of this set of time series in fig.1:

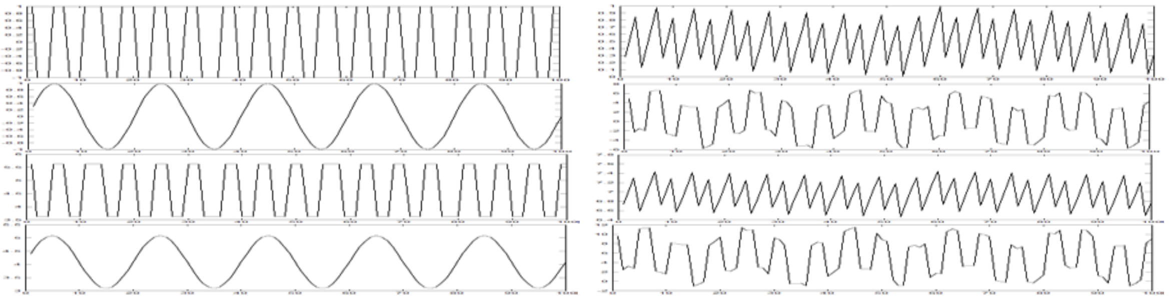

Fig.1: from top to bottom, left to right, the image is y1 to y8 successively.

Follow the steps of principal component analysis, firstly solving the covariance matrix of time series dataset, and secondly obtaining covariance matrix's eigenvalue and eigenvectors (coefficient matrix), its coefficient matrix calculated as follows:

$$T_3 = \begin{pmatrix} 0.1530 & -0.3116 & -0.1260 \\ 0.0029 & 0.0139 & 0.6907 \\ 0.0372 & 0.6336 & -0.0788 \\ 0.6893 & 0.0349 & 0.0293 \\ 0.1530 & -0.3116 & -0.1260 \\ 0.0029 & 0.0139 & 0.6907 \\ 0.0372 & 0.6336 & -0.0788 \\ 0.6893 & 0.0349 & 0.0293 \end{pmatrix}$$

Because of the simulation data designed has less dimension , it can be seen by observation that the first row and the fifth row, second row and the sixth row, the third row and the seventh row, the fourth row and the eighth row is identical Class. This is consistent with the previous hypothesis. Thereby verify the correctness and effectiveness of the clustering algorithm proposed in paper. When the coefficient matrix has too much dimensions, it is difficult to draw conclusion by observation matrix value, instead of that you can calculate the Euclidean distance of coefficient matrix, through the result of the Euclidean distance matrix can be derived time series dataset's clustering result. By means of the above experiment result that the algorithm can not only obtain the right time series data clustering result, but also can effectively reduce the time complexity of the algorithm.

The advantages of time series clustering based on principal component analysis are: 1) the ability of dimension reduction by ideology of principal component analysis, masterly convert original big dataset's clustering issue to solving the coefficient matrix's Euclidean distance matrix issue. 2) By means of calculating the coefficient matrix of Euclidean distance between vectors, obtained clustering results of the original time series dataset. 3) Algorithm is simple and easy to implement, and can effectively reduce the time complexity of the algorithm.

4. Experiments and Analysis of Result

4.1. Experimental Data

In the experiment, we use annual observation data of 2012 on FY-3 MERSI. Because meteorological observation data are numerical data, and has the same length of different dimensions, it is ideally suitable for using algorithm which is described in this paper to do time series clustering analysis. The time series dataset information is mainly space view observation values, different temperature values on the satellite. Where space view observation time series is divided into 20 channels, there are 7 temperature time series, the information of the dataset is shown in Table 2. Using clustering method referred in this paper to analyze which temperature influence the different channels of space view observation values, the result have significance for meteorological calibration work.

Table 2 Introduction of time series Dataset

Name	Meaning	Length
SV_B1	Space view observation value channel 1	365
SV_B2	Space view observation value channel 2	365
⋮	⋮	⋮
SV_B20	Space view observation value channel 20	365
Bracket1	Bracket 1 temperature on satellite	365
Bracket2	Bracket 2 temperature on satellite	365
BB	Blackbody temperature on satellite	365
Cool1	Cool 1 temperature on satellite	365
Cool2	Cool 2 temperature on satellite	365
KMirror	K mirror device temperature	365
VOC	Visible scanner temperature	365

4.2. Experimental Methods

This paper chooses clustering algorithm based on Euclidean distance which referred in introduction to compare with algorithm proposed in paper, by increasing the dimension of time series dataset to compare the time complexity of two algorithms.

Time series clustering algorithm based on principal component analysis proposed in this paper need to enter two parameters, one is time series dataset $\left\{x_1(t), x_2(t), \cdots, x_m(t)\right\}, t=1,2,\cdots,T$ and another is distance threshold. By means of this method can first get the Euclidean distance matrix between different dimensions of coefficient matrix, than use this distance matrix and the input threshold we can finally obtain the clustering result of raw time series dataset.

4.3. Result and Analysis

(1) With the increasing dimension of time series dataset, compare the time complexity of this two algorithms, result is shown in figure 2: where "□" stand for using method based on Euclidean distance, "∗"stand for using method based on principal component analysis.

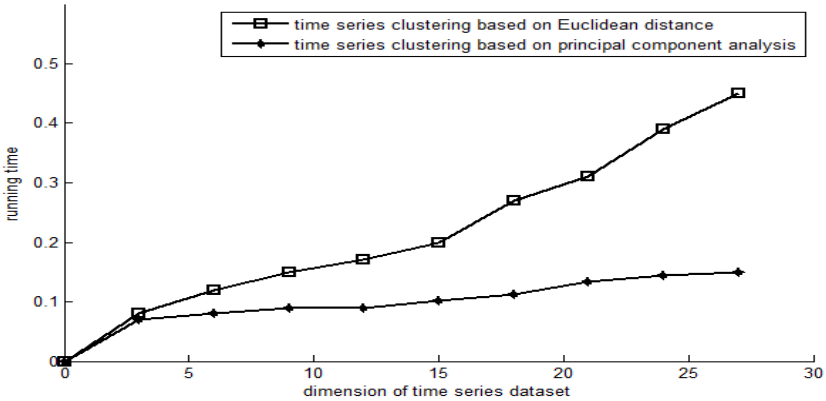

Fig.2 time complexity of two algorithms

It can be seen from Figure 2, with the increasing of data dimension, the running time of algorithm based on PCA is much better than the running time of algorithm based on Euclidean distance. Illustrate that the proposed algorithm in this paper can reduce the dimension of the raw time series dataset, through calculate the coefficient matrix's Euclidean distance to achieve better results.

(2) By using PCA to calculate time series dataset in table 2, the cumulative contribution of the first nine main ingredients in covariance matrix is high, so choose the first nine eigenvalues' corresponding coefficient to do time series clustering, algorithm input threshold is set to 0.1 the results shown in Table 3:

Table 3 clustering result of experiment data

Clustering result	Name of dimension in time series dataset
Set 1	SV_B5, SV_B12, SV_B13, SV_B14, SV_B15, SV_B16, SV_B17, SV_B19, SV_B20,Cool1,Cool2，Kmirror，VOC
Set 2	SV_B8，SV_B9，SV_B10，SV_B11，Bracket1，Bracket2
Set 3	SV_B1，SV_B2，SV_B3，SV_B4，SV_B6，SV_B7，BB

Through the clustering result we can draw conclusion that space view observation value channel 5,12,13,14,15,16,17,19,20 subject to temperature Cool1, Cool2, Kmirror, VOC impact, cluster to a class. Time series observation values of space view channel 8, 9, 10, and 11 influenced by temperature Bracket1, Bracket2, so clustered together. Space view observations channel 1, 2, 3,4,6,7 affected by temperature BB and clustered together. The clustering result takes role as guide for meteorological calibration.

5. Conclusion

There are deficiencies in the time series clustering method based on Euclidean distance, this method is not good at deal with time series dataset which has high dimensions and contain outliers. Combined with idea of principal component analysis in multivariate statistical analysis, first of all find the main ingredient of time series dataset, then solve the coefficient matrix under the main ingredient, through experiments we find that the clustering result of coefficient matrix is consistent with time series dataset. So this article is proposed a time series clustering method based on principal component analysis and Euclidean distance.

Experimental result show that compared with the time series clustering algorithm based on Euclidean distance, the method proposed in this paper can quickly and efficiently obtained the clustering result. Under the same time series dataset, the running time of this method is significantly less than the running time of time series clustering algorithm based on Euclidean distance, and experimental result on simulation data and real data show that it is correct and effective clustering result. Especially when the time series in the dataset has a linear correlation, the approach proposed in this paper has obvious advantages. But the downside of this approach is that when the time series dataset with nonlinear correlation, the result will have mistakes, it needs to continue to improve until achieving better result.

Acknowledgement

This work is supported by Beijing Natural Science Foundation（4144073）.

References

[1] Keogh E,Kasetty S. On the need for time series data mining benchmarks:a survey and empirical demonstration[J].Data Mining and Knowledge Discovery,2003,7(4):349-371

[2] Mahalanobis P C.On the generalized distance in statistics[C]//Proceedings of the National Institute of Sciences of India.1936:49-55.

[3] Berndt D J,Clifford J.Using dynamic time wrapping to find patterns in time series[R].[S.l.]:AAAI94 workshop on knowledge discovery in databases,1994:359-370.

[4] Keogh E J，Chakrabarti K，Pazzani M J，et al.Dimensionality reduction for fast similarity search in large time series databases[J].Journal of Knowledge and Information Systems，2001，3（3）：263-286.

[5] Prat K B , Fink E. Search for patterns in compressed time series[J] . International Journal of Image and Graphics ,2002 , 2 (1) : 89-106.

[6] Lin J,Keogh E,Li W, Experiencing SAX: A novel symbolic representation of time series[J].Data Mining and Knowledge Discovery,2007,15:107-144.

[7] Lin J,Keogh E,Lonardi S,Chiu B. A symbolic representation of time series,with imolications for streaming algorithms[C].Proceeding of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery,San Diego,2003:2-11.

[8] Xiong Y,Yeung D Y.Mixtures of ARMA models for model-based time series clustering[C]//Proceeding of the 2002 IEEE International Conference on Data Mining. Maebaghi City, Japan:IEEE Computer Society,2002:717-720.

[9] T.O,Firoiu L,Cohen P R.Clustering time series with hidden Markov models and dynamic time warping[C]//Proceedings of the IJCAI-99 Workshop on Neural,Symbolic,and Reinforcement Learning Methods for Sequence Learning.Stockholm,Sweden:Morgan Kaufmann,1999.

[10] Keogh E，Folias T.The UCR time series data mining archive[D/OL].Irvine，CA:University of California，Department of Informationand Computer Science，2009.http：//www.cs.ucr，edu/~eamonn/TSDMA/index.html.

[11] LI Zheng-xin, GUO Jian-sheng, HUI Xiao-bin, SONG Fei-fei. Dimension reduction method for multivariate time series based on common principal component[J]. Control and Decision,2013, 1001-0920 (2013) 04-0531-06