



Asynchronism-based principal component analysis for time series data mining



Hailin Li*

College of Business Administration, Huaqiao University, Quanzhou 362021, China

ARTICLE INFO

Keywords:

Asynchronous correlation
Covariance matrix
Principal component analysis
Time series data mining
Dynamic time warping

ABSTRACT

Principal component analysis (PCA) is often applied to dimensionality reduction for time series data mining. However, the principle of PCA is based on the synchronous covariance, which is not very effective in some cases. In this paper, an asynchronism-based principal component analysis (APCA) is proposed to reduce the dimensionality of univariate time series. In the process of APCA, an asynchronous method based on dynamic time warping (DTW) is developed to obtain the interpolated time series which derive from the original ones. The correlation coefficient or covariance between the interpolated time series represents the correlation between the original ones. In this way, a novel and valid principal component analysis based on the asynchronous covariance is achieved to reduce the dimensionality. The results of several experiments demonstrate that the proposed approach APCA outperforms PCA for dimensionality reduction in the field of time series data mining.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Time series is one kind of the most important research objects in the field of data mining. The techniques used in this data is called time series data mining (TSDM) (Esling & Agon, 2012). However, since its high dimensionality often renders standard data mining techniques inefficient, the methods used to reduce the dimensionality are devised.

So far, there exists many methods to resolve this problem, which are considered as two kinds of dimensionality reduction. One is based on univariate time series, such as discrete Fourier transformation (DFT) (Agrawal, Faloutsos, & Swami, 1993), discrete wavelet transformation (DWT) (Maharaj & Urso, 2011; Struzik & Siebes, 1998, 1999), polynomial representation (PR) (Fuchs, Gruber, Pree, & Sick, 2010), piecewise linear approximation (PLA) (Keogh, Chu, Hart, & Pazzani, 2001; Papadakis & Kaburlasos, 2010; Shatkay & Zdonik, 1996), piecewise aggregate approximation (PAA) (Keogh, Chakrabarti, Pazzani, & Mehrotra, 2000; Li & Guo, 2011), symbolic aggregate approximation (SAX) (Lee, Wu, & Lee, 2009; Lin, Keogh, Lonardi, & Chiu, 2003). These methods are mainly proposed to reduced dimensionality from the points of univariate time series. In other word, they mainly concentrate on the transformation of a single time series so that the dimension of the reduced representations is lower than that of the original one. The other is based on the time series dataset, such as singular value decomposition (SVD) (Spiegel, Gaebler, & Lommatzsch,

2011), principal component analysis (PCA) (Singhal & Seborg, 2002) and independent component analysis (ICA) (Cichocki & Amari, 2002).

SVD and PCA are often seen as the same method to retain the first several principal components and to represent the whole dataset. However, ICA is the development of principal component analysis and factor analysis. In the field of time series data mining, the methods are often used and combined with the corresponding measurements to discover the information and knowledge from time series dataset. Krzanowski (1979) used PCA to construct the principal components and chosen the first k principal components to represent the multivariate time series. At the same time, the similarity between two time series are calculated by using the cosine value of the angle between the corresponding principal components. Singhal and Seborg (2005) proposed a new approach S_{dist} to compute the similarity based on PCA, which is better than the earlier methods. Karamitopoulos and Evangelidis (2010) used PCA to construct the feature space of the queried time series and projected every time series to the space. They computed the error between two reconstructed time series as the distance between the query time series and the queried one. SVD is often based on PCA, which uses KL decomposition method to reduce the dimensionality of time series. Li, Khan, and Prabhakaran (2006) proposed two methods to choose the feature vectors and used them to classify time series. Weng and Shen (2008) extended the traditional SVD to an two-dimensional SVD (2dSVD) that extracts the principal components from the column–column and row–row directions to compute the covariance matrix. Since feature extraction is one of the most importance tasks for ICA, it was applied to the analysis of time series. Wu and Yu (2005) used FastICA (Hyvärinen, 1999) to

* Tel.: +86 595 22693815.

E-mail address: hailin@mail.dlut.edu.cn

obtain independent principal components for multivariate time series and cluster them by combining with the correspond distance measurement. Baragona and Battaglia (2007) used ICA to detect the anomalies by extracting the unusual components.

PCA is the basic theory and widely used to reduce the dimensionality of time series (Karamitopoulos & Evangelidis, 2010; Bankó & Abonyi, 2012). It uses the variance to measure how much the information is retained. Moreover, the covariance is computed to measure the correlation between two different time series in PCA. However, the traditional PCA uses the linear and synchronous method to compute the covariance between two time series, which is not effective when the two series are similar or correlative at different points in time. In other words, the same shape trends appearing on two time series at different points in time will be regarded as uncorrelated or negative correlated. It means that in some cases PCA works ineffectively. Moreover, the length of time series must be equal when they are research by PCA. Meanwhile, PCA is often used to mine the knowledge from multivariate time series dataset instead of univariate time series dataset.

The research motivations of this work are to overcome the above mentioned problems. Firstly, the dimensionality of time series with different lengths can be reduced by the principle of principal component. It means that the proposed method can process the time series with different lengths. However, the existing work including SVD, PCA, and ICA only process the ones with equal length. Secondly, the existing methods only consider the synchronous relationship between two variables or two time series, they neglect the asynchronous relationships. Therefore, the proposed methods must be improved to consider the asynchronous relationships. Thirdly, the important information about time series should be concerned by the proposed method rather than the existing work. The reason is that some points of time series reflect the key shape trends and can provide much more important information than the others.

For the above mentioned motivations, the work will include the measurement of asynchronous correlation coefficient, the design of asynchronism-based PCA and the representations of univariate time series for dimensionality reduction. The asynchronous correlation derives from correlation coefficient between a pair of two interpolated time series that are formed by the elements of the best warping path. Moreover, the best warping path can be found by dynamic time warping (DTW) (Yu, Yu, & Hu, 2011). The interpolated time series can be used to improve the effectiveness of correlation coefficient (Pearsons product moment correlation coefficient) (Rodgers & Nicewander, 1988), which measures the similarity (or correlation) between time series with the same shape trends appearing on the different points in time. The asynchronism-based PCA considers the asynchronous correlation to measure the whole time series dataset and obtains the first several principal components that retain the important information about the time series dataset as much as possible. In particular, the tuple of the first several principal components is regarded as the corresponding representations so that every time series can be represented by a short tuple for dimensionality reduction. In comparison to the traditional PCA, the proposed method (Asynchronism-based PCA, APCA) not only can measure the synchronous correlation as PCA does, but also can obtain the asynchronous correlation. It is a good approach to measure the similarity between two time series whose similar shape trends appear on the different points in time.

The remainder of the paper is organized as follows. In Section 2, we provide some necessary background material and discuss related work. In Section 3, we present the proposed method. The experimental evaluation of the new method is described in Section 4. Finally, we discuss our results further and conclude in Section 5.

2. Background and related work

PCA is a common method used to reduce the dimensionality of time series. At the same time, DTW is one of the most robust ways to measure the similarity between time series. According to motivations of this work, the two methods are prior to be introduced in this section.

2.1. Principal component analysis

PCA is a well-known statistical approach that is often used to reduce the dimensionality of dataset (Jolliffe, 2004; Wang, 2012). The dataset can be represented as a data matrix $X_{n \times m}$, where n denotes the number of objects with m properties (or variables). Accordingly, $X_{n \times m}$ denotes a time series dataset that has m time series of length n . Each column represents a time series, and each row represents a group of observed values for a special time.

PCA is an orthogonal linear transformation. It transforms a dataset to a new system. The greatest variance of data points in the new system by any projection of the data objects lies on the first coordinate that we call the first principal component, the second variance on the second coordinate (the second principal component), and so on. In this way, PCA transforms the data matrix X of size $n \times m$ into another reduced matrix Y of size $n \times k$ for the dimensionality reduction, where $k < m$.

Formally, Y is a reduced dataset with k variables (or principal components) that are orthogonal to each other, and $Y_{n \times k} = X_{n \times m} V_{m \times k}$, where $V_{m \times k}$ are composed of the first k principal components and X is zero empirical mean. According to SVD, $\Sigma = V \Lambda V^{-1}$, where Σ is the covariance matrix of X , that is $\Sigma = X^T X$ because of the zero empirical mean. Meanwhile, Λ is an $m \times m$ rectangular diagonal matrix with nonnegative real numbers on the diagonal and the real numbers are composed of the eigenvalues of Σ in descending order. V is the corresponding eigenvector matrix of Σ . Moreover, according to the energy content for each eigenvector, the first k principal components are chosen from the first k eigenvectors of which the cumulative energy content is not less a threshold ε . The algorithm of PCA can be described as follows.

Step 1: Organize the dataset (or data matrix) $X_{n \times m}$. Each row denotes an observation with m variables and each column denotes an variable.

Step 2: Calculate the empirical mean and change the data matrix into a new one with zero mean. At the same time, the original dataset replaces the new one. That is $X = X - SU$, where $U(1, i) = \frac{1}{n} \sum_{j=1}^n X(j, i)$ and $i = 1, 2, \dots, m$ and S is a $n \times 1$ column vector.

Step 3: Compute the eigenvectors and eigenvalues of the covariance matrix $\Sigma = X^T X$. According to SVD, $\Sigma = V \Lambda V^{-1}$, where Λ is the diagonal matrix of eigenvalues of Σ in descending order. Meanwhile, sort the eigenvector matrix V according to the order of decreasing eigenvalues, that is $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_m$.

Step 4: Choose a subset of the eigenvectors as basis vectors according to the cumulative energy content. If the cumulative energy content $g_k = \sum_{i=1}^k \lambda_i$ is above a certain value ε , then the first k eigenvectors are selected as the basis vectors. In other words, if the contribution rate $\eta = \frac{g_k}{g_m}$ of cumulative energy is larger than a threshold ε , then the first k eigenvectors can be seen as the best principal components.

Step 5: Project the original dataset onto the new system. The new dataset Y with low dimension can be formed by $Y_{n \times k} = X_{n \times m} V_{m \times k}$. Since k is often less than m , that is $k < m$, PCA achieves the dimensionality reduction.

In the field of time series data mining, PCA is often used to reduce the dimensionality of multivariate time series and PCA-based measurements are also applied to this field. One of the earliest

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات