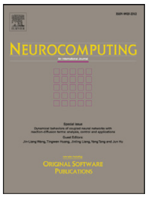
Neurocomputing 273 (2018) 141–151

Contents lists available at ScienceDirect

Neurocomputing

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



Brief papers

Local tangent space alignment via nuclear norm regularization for incomplete data[☆]

Jing Wang[*], Xiaolong Sun, Jixiang Du

The School of Computer Science and Technology, Huaqiao University, Xiamen 361021, PR China

ARTICLE INFO

Article history:
Received 28 December 2016
Revised 28 July 2017
Accepted 30 July 2017
Available online 22 August 2017

Communicated by Huaping Liu

Keywords:
Manifold learning
Incomplete data
Nuclear norm
Regularization

ABSTRACT

Manifold learning approaches seek to find the low-dimensional features of high-dimensional data. When some values of the data are missing, the effectiveness of manifold learning methods may be greatly limited since they have difficulty in determining the local neighborhoods and discovering the local structures of neighborhoods. In this paper, a novel manifold learning approach called local tangent space alignment via nuclear norm regularization (LTSA–NNR) is proposed to discover the nonlinear features of the incomplete data. The neighbors of each sample point are selected using the cosine similarity measurement. A new nuclear norm regularization model is then proposed to discover the local coordinate systems of the determined neighborhoods. Different with the traditional manifold learning approaches, the dimensions of local coordinate systems are various in a reasonable range. The global coordinates of the incomplete data are finally obtained by aligning the local coordinates together. We demonstrate the effectiveness of our method on real-world data sets.

1. Introduction

Machine learning and data mining may involve high-dimensional data sets with arbitrary patterns of missing data in practical applications. For example, in video surveillance applications, part of the target of the monitoring may be in the shade of other objects. The occluded images can be viewed as incomplete data. How to discover the intrinsic structure of incomplete data is increasingly becoming a focus [19,23,25].

In recent years, matrix factorization(MF) approaches are proposed to discover the linear features of incomplete data. They try to find a good approximation to the incomplete data by the product of two or three matrix factors. In general, the regularized techniques are widely used to avoid over-fitting problem [7,8,13]. Some constraints may be also imposed on the matrix factors such as orthogonal constraints[11,14] and non-negative constraints [6,16,24]. Although these approaches have been widely used in some applications such as recommendation system [4,8,13], gene expression analysis [16] and object recognition [6,11], the effectiveness of them may be very limited when the data points lie on or close to a nonlinear manifold.

Recently, manifold learning methods are proposed to learn the nonlinear low-dimensional manifolds from sample data points embedded in high-dimensional spaces. The proposed algorithms include isometric mapping (ISOMAP) [17], locally linear embedding (LLE) [10], laplacian eigenmaps (LE) [1], nonlinear embedding preserving multiple local-linearities (NEML) [22] and local tangent space alignment (LTSA) [29], etc. One basic idea of most manifold learning algorithm is to construct the local neighborhoods of the manifold and linearly approximate the local manifold geometries within the neighborhoods, and then nonlinearly map the points to a lower dimensional space preserving the discovered local geometries. These algorithms have been successfully applied in many fields of information processing due to the simple geometric intuitions, straightforward implementation and global optimization [12,18]. However, they may fail on the data with missing values. For the incomplete data, it is difficult to select the neighbors of each sample point which can reflect the local geometric structure of the manifold. More important, the local geometric structures may be blurry. And it is difficult to exploit the local geometries in the presence of missing values.

More recently, there are some efforts on developing new manifold learning algorithms for incomplete data. Wang et al. proposed a denoising algorithm to reconstruct the missing values of the incomplete data which generalizes matrix completion to curved manifolds [20]. It can be used as a post-processing step on an initial reconstruction of incomplete data, but does not learn the nonlinear features explicitly. In [2], it learns the nonlinear

[☆] The work of this author was supported in part by NSFC (61370006, 61673186), NSF of Fujian Province 2014J01237, and Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University ZQN-PY116.

[*] Corresponding author.
E-mail address: wroaring@hqu.edu.cn (J. Wang).

http://dx.doi.org/10.1016/j.neucom.2017.07.055
0925-2312/© 2017 Elsevier B.V. All rights reserved.

low-dimensional representation of the incomplete data and recoveries the missing values by unsupervised regression approach. However, local optimum exists and good initializations of missing values are also required for the proposed algorithm. In [3], an improved LTSA algorithm called EM-LTSA is proposed to learn manifold from corrupted image set. It obtains the local coordinates using an extended EM-based PCA algorithm instead of standard SVD technique. However, it may have a limited effectiveness on the incomplete image set if there are a considerable number of missing pixels.

Most of the existing algorithms reconstruct the incomplete data before learning the low-dimensional features. The effectiveness of these algorithms highly relies on the accuracy of recovering the missing values. In this paper, we aim to propose an effective manifold learning method to discover the nonlinear features of the incomplete data explicitly without recovering the missing values. The neighborhoods of each sample point are firstly determined using the cosine similarity measurement instead of the Euclidean distance measurement. A new nuclear norm regularization model is then proposed to discover the local coordinate systems of the determined neighborhoods only using the known values. Unlike those work in [3,29], the dimensions of the discovered local coordinate systems are various in a reasonable range rather than a constant value. The various dimensional local coordinate systems are more stable for incomplete data. A novel algorithm called local tangent space alignment via nuclear norm regularization (LTSA–NNR) is then proposed to obtain the global coordinates of the incomplete data by aligning the discovered local coordinates.

The rest of the paper is organized as follows. We briefly review the LTSA algorithm in Section 2 and discuss its failure mode on incomplete data. Local linear fitting via nuclear norm regularization to obtain the local coordinates of incomplete data is proposed in Section 3. Then we propose the main algorithm LTSA–NNR in Section 4. In Section 5, we compare the proposed LTSA–NNR with the original LTSA and its extensions. In Section 6, several experiments are carried out to evaluate LTSA–NNR. Finally, the conclusions are given in Section 7.

2. A brief review of LTSA

The manifold learning approaches can be classified into global methods such as ISOMAP [17] and local methods such as LLE [10] and its variations, LE [1], NEML [22] and LTSA [29], etc. Most of the manifold learning approaches consists of three steps :(1) construct the local neighborhoods, (2) extract the local geometries within the neighborhoods, and (3) minimize a global cost function to obtain the embedding results. In general, the manifold learning approaches are different in the second and third steps. For example, ISOMAP computes the geodesic distances between points, and then projects the points into the low-dimensional space to preserve the geodesic distances. LLE computes the reconstruction weights between each sample point and its neighbors, and then finds the embedding coordinates that are reconstructed by the same weights. LE constructs a weighted graph to exploit the local geometry, and determines a low-dimensional embedding by forcing the neighbors to be close in the embedding space. NEML discovers the local structure of the local neighborhood using multiple linear independently weight vectors, and find the global coordinates in the embedding space to preserve the discovered multiple local-linearities. LTSA constructs local linear fitting to approximate the tangent space at each sample point, and then aligns the local coordinates to obtain the embedding results. Since our work is an extension of LTSA, we outline the basic steps of LTSA as follows.

Given a data set $X = [x_1, \dots, x_N]$ with $x_i \in R^m$, sampled from a d-dimensional manifold. Assuming that d is known, LTSA aims to find the global coordinates $t_1, \dots, t_N$ by proceeding in the following steps:

(1) *Setting local neighborhoods*. For each $x_i, i = 1, \dots, N$, finding its k nearest neighbors (including x_i itself) using Euclidean distance measurement. Denote the neighbor set $X_i = [x_{i_1}, \dots, x_{i_k}]$.

(2) *Extracting local coordinates*. For each $x_i, i = 1, \dots, N$, applying PCA to the neighbor set X_i to approximate the local tangent space of x_i and obtain the local coordinates Θ_i. It can be done by an optimal linear fitting to the sample points in the neighborhood, i.e.,

$$\begin{aligned} &\min_{c,U,\Theta} \|X_i - ce_k^T - U\Theta\|_F^2 \\ &\text{s.t. } c \in R^m, U \in R^{m\times d}, U^TU = I_d, \Theta \in R^{d\times k}, \end{aligned} \tag{1}$$

where $e_k \in R^k$ is a vector of all ones and I_d is a $d \times d$ elementary matrix.

(3) *Aligning local coordinates*. For the local neighborhood X_i, the local reconstruction error

$$\min_{c\in R^d, L\in R^{d\times d}} \|T_i - ce_k^T - L\Theta_i\|_F^2$$

measures the difference between $T_i = [t_{i_1}, \dots, t_{i_k}]$ and the local coordinates Θ_i under the optimal affine transformation. To align the N sets of the local coordinates Θ_i, the global coordinates $T = [t_1, \dots, t_N]$ should minimize the local reconstruction error over all local neighborhoods, i.e., minimize the global reconstruction error

$$\min_{T, T^TT = I_d} \sum_{i=1}^{N} \min_{c\in R^d, L\in R^{d\times d}} \|T_i - ce_k^T - L\Theta_i\|_F^2. \tag{2}$$

The theory analysis about the alignment procedure can be found in [26].

The first row of Fig. 1 illustrates these three steps of LTSA. As it can be seen, the local tangent space can be well approximated by applying PCA to the neighbor set.[1] In the step of alignment, the local coordinate system in the embedding space can match the discovered local coordinate system (the green star points) well. And the embedding results show that LTSA can recover the arc length of the curve with an affine transformation. Generally, LTSA can work well on the data which are well sampled from the manifold. However, LTSA may fail on the incomplete data. The data point with missing values acts as an outlier. The second row of Fig. 1 illustrates the failure of LTSA on the incomplete data, due to the following reasons: (1) The neighbors of the incomplete points may be wrongly determined using the Euclidean distance measurement. (2) The local PCA cannot give an acceptable estimate to the tangent space at the incomplete point, and the local coordinates have large deviations to the true coordinates.

The example in Fig. 1 clearly shows that it is desirable to have better strategies for determining the local neighborhoods and extracting the local coordinates. In the paper, our algorithm aims to find the correct neighbors only using the known values and extract the true coordinates of the local neighborhoods of the incomplete points. See the third row of Fig. 1 for illustrating the motivation of our algorithm.

3. Local linear fitting via nuclear norm regularization

Assume that we are given a set of incomplete data points $x_1, \dots, x_N$, where $x_i \in R^m$ with missing values are sampled from a d-dimensional manifold. We represent the pattern of missing values in x_i with an indication vector $f_i = [f_{i1}, f_{i2}, \dots, f_{im}]^T$ where

[1] The angle between the local tangent space (red line) and span(U_i) (blue line) approximates to 0.