Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Pattern Recognition 39 (2006) 1053 – 1065

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Robust locally linear embedding

Hong Chang, Dit-Yan Yeung*

Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

Received 20 August 2004; received in revised form 11 May 2005; accepted 12 July 2005

Abstract

In the past few years, some nonlinear dimensionality reduction (NLDR) or nonlinear manifold learning methods have aroused a great deal of interest in the machine learning community. These methods are promising in that they can automatically discover the low-dimensional nonlinear manifold in a high-dimensional data space and then embed the data points into a low-dimensional embedding space, using tractable linear algebraic techniques that are easy to implement and are not prone to local minima. Despite their appealing properties, these NLDR methods are not robust against outliers in the data, yet so far very little has been done to address the robustness problem. In this paper, we address this problem in the context of an NLDR method called locally linear embedding (LLE). Based on robust estimation techniques, we propose an approach to make LLE more robust. We refer to this approach as robust locally linear embedding (RLLE). We also present several specific methods for realizing this general RLLE approach. Experimental results on both synthetic and real-world data show that RLLE is very robust against outliers.

© 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Nonlinear dimensionality reduction; Manifold learning; Locally linear embedding; Principal component analysis; Outlier; Robust statistics; M-estimation; Handwritten digit; Wood texture

1. Introduction

Dimensionality reduction is concerned with the problem of mapping data points that lie on or near a low-dimensional manifold in a high-dimensional data space to a low-dimensional embedding space. Traditional techniques such as principal component analysis (PCA) and multidimensional scaling (MDS) have been extensively used for linear dimensionality reduction. However, these methods are inadequate for embedding nonlinear manifolds.

In recent years, some newly proposed methods such as isometric feature mapping (Isomap) [1], locally linear embedding (LLE) [2,3], and Laplacian eigenmap [4,5] have aroused a great deal of interest in nonlinear dimensionality reduction (NLDR) or nonlinear manifold learning problems. Unlike previously proposed NLDR methods such as autoassociative neural networks which require complex optimization techniques, these new NLDR methods enjoy the primary advantages of PCA and MDS in that they still make use of simple linear algebraic techniques that are easy to implement and are not prone to local minima.

Despite the appealing properties of these new NLDR methods, they are not robust against outliers in the data. Although some extensions have been proposed to the original methods [6–12,3,13–15], very little has yet been done to address the outlier problem. Among the extensions proposed is an interesting extension of LLE proposed by Teh and Roweis, called locally linear coordination (LLC) [13], which combines the subspace mixture modeling approach with LLE. A recent work by de Ridder and Franc [16] attempted to address the outlier problem by proposing a robust version of LLC based on a recent development in the statistics community called mixtures of t-distributions. However, although the robust version of LLC is less sensitive to outliers than LLC, the authors found that it is still more sensitive to outliers than ordinary LLE. Zhang and Zha [17] proposed a preprocessing method for outlier removal and noise reduction before NLDR is performed. It is based on a weighted version of PCA. However, the method for determining the weights is heuristic in nature without formal justification. More recently, Hadid and Pietikäinen [18] studied

* Corresponding author. Tel.: +852 2358 6977; fax: +852 2358 1477.
E-mail address: dyyeung@cs.ust.hk (D.-Y. Yeung).

0031-3203/$30.00 © 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2005.07.011

the outlier problem and proposed a method to make LLE more robust. However, their method is also heuristic in nature. Moreover, their method is based on the assumption that all outliers are very far away from the data on the manifold and they themselves form distinct connected components in the neighborhood graph. Hence the outliers have no effect on the reconstruction of the manifold data points. Apparently, this assumption is not always true for many real-world applications.

In this paper, we address the outlier problem in the context of LLE. Based on robust PCA techniques, we propose an approach to make LLE more robust. The rest of this paper is organized as follows. In Section 2, we first give a quick review of the LLE algorithm. In Section 3, the sensitivity of LLE to outliers is illustrated through some examples based on synthetic data. A new approach called robust locally linear embedding (RLLE) is then presented in Section 4 together with several specific realizations of the approach. Section 5 shows some experimental results to demonstrate the effectiveness of RLLE in the presence of outliers. Some concluding remarks are given in Section 6.

2. Locally linear embedding

Let $\mathscr{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a set of N points in a high-dimensional data space $\mathscr{R}^D$. The data points are assumed to lie on or near a nonlinear manifold of intrinsic dimensionality $d < D$ (typically $d \ll D$). Provided that sufficient data are available by sampling well from the manifold, the goal of LLE is to find a low-dimensional embedding of $\mathscr{X}$ by mapping the D-dimensional data into a single global coordinate system in $\mathscr{R}^d$. Let us denote the corresponding set of N points in the embedding space $\mathscr{R}^d$ by $\mathscr{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$.

The LLE algorithm [3] can be summarized as follows:

(1) For each data point $\mathbf{x}_i \in \mathscr{X}$:

   (a) Find the set $\mathscr{N}_i$ of K nearest neighbors of $\mathbf{x}_i$.

   (b) Compute the reconstruction weights of the neighbors that minimize the error of reconstructing $\mathbf{x}_i$.

(2) Compute the low-dimensional embedding $\mathscr{Y}$ for $\mathscr{X}$ that best preserves the local geometry represented by the reconstruction weights.

Step (1)(a) is typically done by using Euclidean distance to define neighborhood, although more sophisticated criteria may also be used.

Based on the K nearest neighbors identified, step (1)(b) seeks to find the best reconstruction weights. Optimality is achieved by minimizing the local reconstruction error for $\mathbf{x}_i$

$$\mathscr{E}_i = \left\| \mathbf{x}_i - \sum_{\mathbf{x}_j \in \mathscr{N}_i} w_{ij} \mathbf{x}_j \right\|^2, \tag{1}$$

which is the squared distance between $\mathbf{x}_i$ and its reconstruction, subject to the constraints $\sum_{\mathbf{x}_j \in \mathscr{N}_i} w_{ij} = 1$ and $w_{ij} = 0$ for any $\mathbf{x}_j \notin \mathscr{N}_i$. Minimizing $\mathscr{E}_i$ subject to the constraints is a constrained least squares problem. After repeating steps (1)(a) and (1)(b) for all N data points in $\mathscr{X}$, the reconstruction weights obtained form a weight matrix $\mathbf{W} = [w_{ij}]_{N \times N}$.

Step (2) of the LLE algorithm is to compute the best low-dimensional embedding $\mathscr{Y}$ based on the weight matrix $\mathbf{W}$ obtained. This corresponds to minimizing the following cost function:

$$\Phi = \sum_{i=1}^{N} \left\| \mathbf{y}_i - \sum_{\mathbf{x}_j \in \mathscr{N}_i} w_{ij} \mathbf{y}_j \right\|^2, \tag{2}$$

subject to the constraints $\sum_{i=1}^{N} \mathbf{y}_i = \mathbf{0}$ and $1/N \sum_{i=1}^{N} \mathbf{y}_i \mathbf{y}_i^{\mathrm{T}} = \mathbf{I}$, where $\mathbf{0}$ is a column vector of zeros and $\mathbf{I}$ is an identity matrix. Note the similarity of this equation to (1). Based on $\mathbf{W}$, we can define a sparse, symmetric, and positive semidefinite matrix $\mathbf{M}$ as follows:

$$\mathbf{M} = (\mathbf{I} - \mathbf{W})^{\mathrm{T}} (\mathbf{I} - \mathbf{W}).$$

Note that (2) can be expressed in the quadratic form, $\Phi = \sum_{i,j} M_{ij} \mathbf{y}_i^{\mathrm{T}} \mathbf{y}_j$, based on $\mathbf{M} = [M_{ij}]_{N \times N}$. By the Rayleigh–Ritz theorem [19], minimizing (2) with respect to the $\mathbf{y}_i$'s in $\mathscr{Y}$ can be done by finding the eigenvectors with the smallest (nonzero) eigenvalues.

Fig. 1 shows how LLE works in finding the low-dimensional embedding of the S curve manifold from $\mathscr{R}^3$ to $\mathscr{R}^2$.

3. Sensitivity of locally linear embedding to outliers

In this section, we will show through examples how the LLE results can be affected by outliers in the data. We use three artificial data sets that have been commonly used by other researchers: Swiss roll (Fig. 2), S curve (Fig. 3), and helix (Fig. 4). For each data set, uniformly distributed random noise points that are at least at a certain distance from the data points on the manifold are added as outliers. Table 1 shows the parameter settings used in these experiments. The parameters include the dimensionality of the data space D, the dimensionality of the embedding space d (i.e., intrinsic dimensionality of the nonlinear manifold), the number of nearest neighbors K, the number of clean data points on the manifold, the number of outlier points, and the minimum distance between randomly generated outliers and data points on the manifold.

As we can see from subfigures (b) of Figs. 2–4, LLE cannot preserve well the local geometry of the data manifolds in the embedding space when there are outliers in the data. In fact, in the presence of outliers, the K nearest neighbors of a (clean) data point on the manifold may no longer lie on a locally linear patch of the manifold, leading to a small bias to the reconstruction. As for an outlier point, its neigh-

Download English Version:

https://daneshyari.com/en/article/531997

Download Persian Version:

https://daneshyari.com/article/531997

Daneshyari.com