Neurocomputing 110 (2013) 153–159

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Letters

Normalized discriminant analysis for dimensionality reduction

Zhizheng Liang*, Shixiong Xia, Yong Zhou

School of Computer Science and technology, China University of Mining and Technology, China

ARTICLE INFO

Article history:
Received 30 April 2012
Received in revised form
2 December 2012
Accepted 8 December 2012
Available online 31 December 2012

Keywords:
Dimensionality reduction
Normalized Laplacain matrix
Locality preserving projections (LPP)
Normalized discriminant analysis (NDA)
Data classification

ABSTRACT

In this paper, we propose the normalized discriminant analysis (NDA) technique for dimensionality reduction. NDA is built on the information of data point pairs that is implicitly encoded by using the pseudo-Riemannian metric tensor. This makes NDA to be easily adapted for unsupervised or supervised learning. It is also interesting to note that the solution of NDA will asymptotically converge to that of generalized linear discriminant analysis (GLDA) under proper conditions. This gives us some insights in understanding the evolving behavior of NDA. Extensive experiments on a simulated data, face images, character images, and UCI data sets are carried out to demonstrate the effectiveness of NDA.

1. Introduction

Dimensionality reduction techniques have been widely used in various areas such as machine learning, pattern recognition, and computer vision [1–3]. The general aim of dimensionality reduction is to reduce the dimensionality of data such that the extracted features are as representable as possible. During the past several decades, a variety of algorithms and techniques [4–7] for dimensionality reduction have been developed. Among them, principal component analysis (PCA) and linear discriminant analysis (LDA) are two widely used linear dimensionality reduction methods. In general, PCA is to obtain an orthogonal set of vectors by maximizing the variance of the projected vectors while LDA is to search for discriminant vectors such that the ratio of the between-class distance to the within-class distance is maximized. However, LDA often suffers from the small sample size (3S) problem if the dimension of data is much larger than the number of data points. To overcome this problem, some effective approaches have been proposed. The algorithms such as regularized LDA, PCA plus LDA [2], pseudo-inverse LDA, and orthogonal LDA [8] mainly handle the singularity problem in LDA. LDA/QR [6] and the spectral regression discriminant analysis (SRDA) [9–11] not only overcome the singularity problem in LDA but also have clearly computational advantages over most LDA algorithms on the large data sets.

Although PCA and LDA can effectively extract the features of data, they are in nature linear and may fail to reveal the underlying structure of some complex data such as faces and handwritten characters. To capture the geometrical structure of data, some non-linear dimensionality reduction algorithms such as ISOMAP [12], LLE [13] and Laplacian Eigenmaps [14] have been developed. Note that these non-linear dimensionality reduction algorithms are only available for the samples in the training set. In order to make these algorithms be adapted for pattern recognition tasks such as face recognition, one often obtains linear approximations of these non-linear dimensionality reduction methods. For example, the locality preserving projections (LPP) [15] algorithm, as a linear approximation of Laplacian Eigenmaps, is an effective dimensionality reduction technique and has many similar properties with Laplacian Eigenmaps.

In this paper, motivated by the normalized Laplacian matrix in graph theory, we propose normalized discriminant analysis for dimensionality reduction. First, we define a normalized total scatter matrix, which extends the total scatter matrix. Then we define a normalized within-locality scatter matrix from the idea of locality preserving projections and further obtain a new scatter matrix. Based on these three scatter matrices, we define two criteria to obtain the projection matrix: one is similar to the generalized Fisher criterion with a Rayleigh quotient in form and the other is similar to the discrepancy criterion. Note that optimizing these two criteria can be transformed into solving the (generalized) eigenvalue problems. Different from classical LDA, normalized discriminant analysis is available for unsupervised or supervised learning. Moreover, it is found that classical linear discriminant analysis is an extreme case of normalized discriminant analysis under proper conditions.

2. Locality preserving projections

The locality preserving projections (LPP) algorithm [15], as a linear approximation of the non-linear Laplacian Eigenmap, is

* Corresponding author.
E-mail address: cuhk_liang@yahoo.cn (Z. Liang).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.12.007

justified to preserve the neighborhood structure in a certain sense. To be specific, LPP searches for an $n \times d$ matrix $\boldsymbol{G}$ to project l samples $\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_l \in \Re^n$ into a low-dimensional subspace in which the local structure of data can be preserved. The linear transformation $\boldsymbol{G}$ is usually obtained by minimizing the following objective function under approximate constraints

$$\min_{\boldsymbol{G}} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \|\boldsymbol{G}^T\mathbf{x}_i-\boldsymbol{G}^T x_j\|^2 S(i,j), \tag{1}$$

where $S(i,j)$is the weight of an edge between the nodes i and j in the adjacency graph. In an adjacency graph, there exists the edge between the nodes i and j if these two nodes belong to k-nearest neighbors or ε-neighborhoods. The weight of an edge between nodes i and j may take the heat kernel $S(i,j)=e^{-\|\mathbf{x}_i-\mathbf{x}_j\|^2/\delta}$or 1 in terms of the connected conditions.

The minimization problem of Eq. (1) along with a constraint condition of vertices constructs the following generalized eigenvalue problem.

$$\boldsymbol{XLX}^T\boldsymbol{g}_k = \lambda_k \boldsymbol{XD}_1\mathbf{X}^T\boldsymbol{g}_k, \quad k=1,\cdots,d, \tag{2}$$

where $\mathbf{D}_1$ is a diagonal matrix whose ith diagonal element is$D_1(i,i)=\sum_j S(i,j)$, $\mathbf{L}=\mathbf{D}_1-\mathbf{S}$, and $\mathbf{g}_k$is the kth column of the matrix $\boldsymbol{G}$. Note that directly solving Eq. (2) may not be efficient on large data sets. To this end, Cai et al. [16] transformed LPP into a regression framework. In this framework, the spectral regression locality preserving projection (SRLPP) can be solved by a sparse matrix eigen-decomposition followed with regularized least squares.

3. Normalized discriminant analysis

In this section, we first give the definition of the normalized total scatter matrix, which generalizes the total scatter matrix in some sense. Then we give two new scatter matrices. Finally, we construct the optimization models and give their algorithm.

Given a random variable $\boldsymbol{x}$, the total scatter matrix is defined as

$$\mathbf{S}_t = E(\boldsymbol{x}-E\boldsymbol{x})(\boldsymbol{x}-E\boldsymbol{x})^T, \tag{3}$$

where $E(\boldsymbol{x})$ denotes the expectation of a random variable $\boldsymbol{x}$. When a set of samples $(\mathbf{x}_1,\cdots,\mathbf{x}_l)$are given, the total scatter matrix is approximately obtained by

$$\tilde{\mathbf{S}}_t = \frac{1}{l}\sum_{i=1}^{l}(\mathbf{x}_i-\mathbf{m})(\mathbf{x}_i-\mathbf{m})^T, \tag{4}$$

where $\boldsymbol{m}=(1/l)\sum_{i=1}^{l}\boldsymbol{x}_i$is the mean of the samples. Note that $\tilde{\mathbf{S}}_t$ can be written as

$$\begin{aligned}\tilde{\boldsymbol{S}}_t &= \frac{1}{l^2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\boldsymbol{x}_i-\boldsymbol{m})(\boldsymbol{x}_i-\boldsymbol{m})^T \\ &= \frac{1}{2l^2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\boldsymbol{x}_i-\boldsymbol{m})(\boldsymbol{x}_i-\boldsymbol{m})^T \\ &\quad + \frac{1}{2l^2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\boldsymbol{x}_j-\boldsymbol{m})(\boldsymbol{x}_j-\boldsymbol{m})^T.\end{aligned} \tag{5}$$

It is not difficult to verify that $\sum_{i=1}^{l}\sum_{j=1}^{l}(\boldsymbol{x}_i-\boldsymbol{m})(\boldsymbol{m}-\boldsymbol{x}_j)^T=0$ and $\sum_{i=1}^{l}\sum_{j=1}^{l}(\boldsymbol{m}-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{m})^T=0$.Adding two expressions to Eq. (5), one can obtain

$$\begin{aligned}\tilde{\boldsymbol{S}}_t &= \frac{1}{2l^2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\boldsymbol{x}_i-\boldsymbol{m}+\boldsymbol{m}-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{m}+\boldsymbol{m}-\boldsymbol{x}_j)^T \\ &= \frac{1}{2l}\sum_{i=1}^{l}\sum_{j=1}^{l}\left(\frac{\boldsymbol{x}_i}{\sqrt{l}}-\frac{\boldsymbol{x}_j}{\sqrt{l}}\right)\left(\frac{\boldsymbol{x}_i}{\sqrt{l}}-\frac{\boldsymbol{x}_j}{\sqrt{l}}\right)^T.\end{aligned} \tag{6}$$

From Eq. (6), it is found that $\tilde{\mathbf{S}}_t$ is expressed in terms of data point pairs. From a viewpoint of graph theory, the data points can be considered as the vertices in an undirected graph where the weight between data points $\mathbf{x}_i$ and $\mathbf{x}_j$ is 1, and the degree of each vertex is l. Thus $\tilde{\mathbf{S}}_t$ can be obtained from a fully connected graph. According to Eq. (6), we define the following normalized total scatter matrix

$$\begin{aligned}\tilde{\mathbf{S}}_{nt} &= \frac{1}{2l}\sum_{i=1}^{l}\sum_{j=1}^{l}W(i,j)\left(\frac{\mathbf{x}_i}{\sqrt{d_i}}-\frac{\mathbf{x}_j}{\sqrt{d_j}}\right)\left(\frac{\mathbf{x}_i}{\sqrt{d_i}}-\frac{\mathbf{x}_j}{\sqrt{d_j}}\right)^T \\ &= \frac{1}{l}\boldsymbol{X}\left(\boldsymbol{I}_{l\times l}-\boldsymbol{D}_2^{-1/2}\boldsymbol{W}\boldsymbol{D}_2^{-1/2}\right)\boldsymbol{X}^T,\end{aligned} \tag{7}$$

where $W(i,j)$ is the weight of data points $\mathbf{x}_i$ and $\mathbf{x}_j$, d_j is the row sum of the matrix $\mathbf{W}$, $\mathbf{D}_2=diag(d_1,\cdots,d_l)$, $\mathbf{I}_{l\times l}$is an $l\times l$ identity matrix, and $\mathbf{D}_2^{-1/2}$is the negative square root of $\mathbf{D}_2$. It is clear that d_j is the degree of data point $\mathbf{x}_j$ and $\mathbf{I}_{l\times l}-\mathbf{D}_2^{-1/2}\mathbf{W}\mathbf{D}_2^{-1/2}$ is the normalized Laplacian matrix of $\boldsymbol{W}$ [17]. Note that the definition of the normalized scatter matrix in Eq. (7) is different from that of the weighted scatter matrix. It is observed that Eq. (7) not only considers the weight of data points but also uses the distribution of data points. Specifically speaking, each data point is also assigned to an additional weight. If the heat kernel [14] is chosen as the weight for Eq. (7), one can see that the following proposition holds in terms of the law of large numbers in probability theory.

Proposition 1. For any positive number ε, one has

$$\lim_{l\to\infty,\delta\to\infty} Prob\left(\|\tilde{\mathbf{S}}_{nt}-\mathbf{S}_t\|_F<\varepsilon\right)=1$$

The proposition shows that $\tilde{\mathbf{S}}_{nt}$will approach $\mathbf{S}_t$ in probability if the parameter of heat kernels and the number of samples both approach the infinity. In real applications, we often need to estimate $\mathbf{S}_t$ on the limited samples. Since $\tilde{\mathbf{S}}_{nt}$will asymptotically converge to $\mathbf{S}_t$ under proper conditions, $\tilde{\mathbf{S}}_{nt}$ can be used as a rough estimation of $\mathbf{S}_t$. Note also that $\tilde{\mathbf{S}}_{nt}$will approach $\tilde{\mathbf{S}}_t$ on the limited samples as the parameter of the heat kernel approaches infinity. To be specific, $\tilde{\mathbf{S}}_{nt}$ extends $\tilde{\mathbf{S}}_t$ in some sense, i.e., $\tilde{\mathbf{S}}_t$ is an extreme case of $\tilde{\mathbf{S}}_{nt}$. Thus one can explore the behavior of the neighborhood of $\tilde{\mathbf{S}}_t$ by changing the parameter of the heat kernel in $\tilde{\mathbf{S}}_{nt}$. In addition, it is found that all the training samples lead to the common vector when they are projected into the null space of $\tilde{\mathbf{S}}_t$ [8]. However, it is seen from Eq. (7) that all the samples may yield different vectors in the null space of $\tilde{\mathbf{S}}_{nt}$ since the degree of each vertex may be different. This shows that the samples also contain discriminant information even if they are projected into the null space of $\tilde{\mathbf{S}}_{nt}$in the general case. For the sake of notational simplicity, we refer to the null space of $\tilde{\mathbf{S}}_{nt}$ as a trivial space.

Given a set of data points, some of them have similar properties. In order to explore the structure of data points with similar properties, we define the following normalized within-locality scatter matrix from the idea of locality preserving projections [15].

$$\begin{aligned}\tilde{\boldsymbol{S}}_{nw} &= \frac{1}{2l}\sum_{i=1}^{l}\sum_{j=1}^{l}S(i,j)\left(\frac{\boldsymbol{x}_i}{\sqrt{s_i}}-\frac{\boldsymbol{x}_j}{\sqrt{s_j}}\right)\left(\frac{\boldsymbol{x}_i}{\sqrt{s_i}}-\frac{\boldsymbol{x}_j}{\sqrt{s_j}}\right)^T \\ &= \frac{1}{l}\boldsymbol{X}\left(\boldsymbol{I}_{l\times l}-\boldsymbol{D}_3^{-1/2}\boldsymbol{S}\boldsymbol{D}_3^{-1/2}\right)\boldsymbol{X}^T,\end{aligned} \tag{8}$$

where $\mathbf{D}_3$ is a diagonal matrix whose diagonal elements are $s_1,\ldots,s_l$, s_i is the degree of data point $\mathbf{x}_i$, and $\mathbf{I}_{l\times l}-\mathbf{D}_3^{-1/2}S\mathbf{D}_3^{-1/2}$ is the normalized Laplacian matrix of $\boldsymbol{S}$ [17]. In Eq. (8), it is required that the data point which pairs with similar properties should have non-zero non-negative weights while the data point pairs with dissimilar properties should have zero weights. Thus Eq. (8) only considers data point pairs with similar properties. It is clear that the main difference between Eq. (8) and Eq. (7) is that the former is defined in terms of the adjacency graph and the latter is constructed based on a fully connected graph where there have the edges for any pair of nodes. If $s_i(i=1,\cdots,l)$ in Eq. (8) are set to 1, we refer to Eq. (8) as the unnormalized within-locality scatter matrix (UWLSM). It is observed that all the training samples with