Neurocomputing 74 (2011) 1478–1484

Contents lists available at ScienceDirect

Neurocomputing

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

Letters

Kernel-view based discriminant approach for embedded feature extraction in high-dimensional space

Miao Cheng [a,b,*], Bin Fang [a], Chi-Man Pun [b], Yuan Yan Tang [a,b]

[a] Department of Computer Science, Chongqing University, Chongqing, China
[b] Department of Computer and Information Science, University of Macau, Macau

ARTICLE INFO

Article history:
Received 7 January 2010
Received in revised form
21 December 2010
Accepted 2 January 2011
Communicated by K. Li
Available online 24 February 2011

Keywords:
Dimensionality reduction
Local discriminant analysis
Curse-of-dimensionality
Kernel analysis
Feature extraction

ABSTRACT

Derived from the traditional manifold learning algorithms, local discriminant analysis methods identify the underlying submanifold structures while employing discriminative information for dimensionality reduction. Mathematically, they can all be unified into a graph embedding framework with different construction criteria. However, such learning algorithms are limited by the curse-of-dimensionality if the original data lie on the high-dimensional manifold. Different from the existing algorithms, we consider the discriminant embedding as a kernel analysis approach in the sample space, and a kernel-view based discriminant method is proposed for the embedded feature extraction, where both PCA pre-processing and the pruning of data can be avoided. Extensive experiments on the high-dimensional data sets show the robustness and outstanding performance of our proposed method.

1. Introduction

It is well known that many machine learning and data mining problems deal with the high-dimensional data representation and analysis. In the past few decades, numerous dimension reduction and feature extraction methods have been devoted to find the resultful feature representation of the original data. In the literature, principal component analysis (PCA) [1] and linear discriminant analysis (LDA) [2,3] have been the most popular techniques. Moreover, they can be carried out in a reproducing kernel Hilbert space (RKHS) by making use of the well-known "kernel trick" [4], and the kernel discriminant method and its variants [5–7] are drawn in the recent years.

Different from the statistical feature extraction techniques that consider the global Euclidean structure, the theoretical basis of manifold learning methods, e.g., ISOMAP [8], LLE [9] and Laplacian eigenmap (LE) [10], depends on the observation that the high-dimensional data may reside on an intrinsic nonlinear manifold with much low dimensionality. Particularly, locality preserving projection (LPP) [11] is presented based on the LE idea. In order to find the discriminative submanifold structures embedded on the original data, some locally supervised learning techniques [12–16] are proposed, in the light of the locality preserving conception.

On the other hand, many real-world applications, such as image retrieval and pattern recognition, handle the high-dimensional data that bring the curse-of-dimensionality for the local discriminant analysis techniques. Generally, the existing local methods handle the high dimensionality of data in the following two ways:

- Since the maximum margin criterion (MMC) has been successfully applied to classical LDA [17], it is employed to replace the ratio discriminant formulation with the subtraction one. The drawback of such an idea is that the computational expense usually depends on the dimensionality of data and it is hard to carry out if the original data lie on a high-dimensional manifold. A simple solution to this problem is to resize the data into a smaller size. Such an idea, however, would destroy the integrity of the original data.
- By other means, PCA is usually used to reduce dimension primarily in such environments. Though the PCA pre-processing step can be considered to generate a new coordinate system, the local manifold structure cannot be preserved if a pruning of PCA energy is adopted.

In view of this intrinsic limitation, we propose a kernel-view based discriminant approach, namely KVDA, for embedded feature extraction of high-dimensional data. Different from other

* Corresponding author at: Department of Computer Science, Chongqing University, Chongqing, China.
E-mail address: mewcheng@gmail.com (M. Cheng).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.01.004

Table 1
Notations.

Notation	Description
X	Original data set
d	Dimensionality of original samples
M, M_p	Scatter matrix
S, S_p	Adjacency matrix
α	Discriminant coefficient
K_p	Penalty kernel matrix
N_p	Amount of graph edges
t	Rank of scatter matrix
$\hat{M}$	Reduced scatter matrix
$\widetilde{M}$,$\widetilde{M}_p$	Kernel scatter matrix
A^T	Transpose of matrix A
n	Amount of samples
K	Inner product kernel matrix
L, L_p	Laplacian matrix
D,D_p,Σ	Diagonal matrix
$L_p{}'$	Indicator matrix
W	Projection matrix
r	Desired features
I	Identity matrix
Λ,Γ	Eigenvalue matrix
V, H	Eigenvector matrix
z	Eigenvector

techniques, it is insensitive to the high dimensionality of data and PCA stage is unnecessary. It is noticeable that though local discriminative learning is conducted via a kernel approach in our work, the original sample space is involved rather than the RKHS.

The remainder of this paper is organized as follows. The problem statement is given in Section 2. The proposed KVDA is described in Section 3. A comprehensive set of comparison experiments on feature extraction and classification is given in Section 4, followed by the conclusion in Section 5. For convenience, the important notations used in the paper are listed in Table 1.

2. Problem statement

To discover the action on the data structure, the difference between PCA and local manifold learning methods for dimensionality reduction is considered. Taking LPP for example, we show the difference between PCA and LPP for the real data sets, namely, Ionosphere and Monks3 [18], in Fig. 1, where the representation of PCA and LPP for the data sets in the first two significant dimensions is illustrated. Evidently, LPP tends to preserve the *local* localities, while PCA makes the *global* distribution of data maximum. In other words, PCA aims to disperse data in a total statistical theory if some data energy is pruned in the reduced subspace, which is contrary to the local ones. It implies that the local structure of data may be destroyed in the predigested PCA subspace.

In the literature, some works have affirmed that LDA can be performed in PCA transformed space in theory [19], and no discriminant power will be lost if the complete data energy is kept. It is well known that the PCA measures the variance of data via a covariance matrix which is equivalent to the total scatter S_t in LDA. Differently, the local discriminant methods are designed under the local graphs based discriminant criterion that have no relationship with the S_t in general. To make the difference clear, the learned subspaces of LDA and the local discriminant method are illustrated in Fig. 2, and the results are computed based upon nine different objects (the top row in Fig. 3) from the ALOI database. In addition, it is noticeable that the most discriminative information included in the range space of the between-class scatter S_b can still be preserved if several principal components are pruned in the PCA stage of LDA, while the discriminant space of local methods can fill the PCA subspace. In other words, pruning of principal components that is widely used in many applications destroys the submanifold structure in the projected subspace. In terms of this, it may be optimal if PCA stage can be abandoned in the local manifold discrimination.

3. Embedded feature extraction under a kernel view

3.1. Kernel-view based discriminant approach

In order to depict the supervised manifold structure hidden in high-dimensional data, two graphs, i.e., the intrinsic graph G and the penalty graph G_p, are generally constructed. Suppose that $X \in \mathbb{R}^{d\times n}$ denotes the data matrix consisting of n samples, S and S_p are respectively the adjacency matrices of the intrinsic and penalty graphs, their corresponding Laplacian matrices are indicated by L and L_p. Without loss of generality, discriminant embedding aims to find an optimal subspace, where the distances of data pairs in G are shortened and the edges in G_p can be enlarged, by optimizing the following objective function:

$$w^* = \arg\min_w \frac{w^T M w}{w^T M_p w} = \arg\min_w \frac{w^T X L X^T w}{w^T X L_p X^T w}. \tag{1}$$

Here, M and M_p indicate the intraclass and extraclass scatter matrices, L and L_p are determined by $L = D - S$ and $L_p = D_p - S_p$, D and D_p are diagonal matrices, of which diagonal entries are column (or row) sum of S and S_p as defined in the graph embedding framework [12]. Then, the optimal w can be obtained by solving a generalized eigenvalue decomposition problem, $Mw = \lambda M_p w$.

Mathematically, the objective function can be optimized via solving the equivalent problem:

$$W^* = \arg\min_W \mathrm{tr}(WMW) \tag{2}$$

subject to

$$W^T M_p W = I, \tag{3}$$

where I indicates the identity matrix. For the high-dimensional data, the range space of M_p that covers the most discriminative features is taken into account. For the sake of simplification, assume that the adjacency weight of the connected intraclass and extraclass edges are all set to be one. Then, the generalized discriminant embedding can be regarded as a kernel approach based on the following proposition [16].

Proposition 1. *Given graph G_p with adjacency matrix $S_p \in \mathbb{R}^{n\times n}$ and Laplacian matrix $L_p \in \mathbb{R}^{n\times n}$, there exists another indicator matrix $L_p' \in \mathbb{R}^{n\times N_p}$ constructed according to the N_p elements that are equal to "1" in the upper (or lower) triangular matrix of S_p. That is, for each "1" in upper (or lower) triangular matrix of S_p, there is a corresponding column in L_p' as*

$$\left(\overbrace{0\ldots0}^{1,\ldots,i-1},\overset{i}{1},\overbrace{0\ldots0}^{i+1,\ldots,j-1},\overset{j}{-1},\overbrace{0\ldots0}^{j+1,\ldots,n}\right)^T \tag{4}$$

Then, L_p can redefined as

$$L_p = L_p' L_p'^T. \tag{5}$$

Let K be the inner product matrix $X^TX \in \mathbb{R}^{n\times n}$, K_p denote the penalty kernel matrix $L_p'^T K L_p' \in \mathbb{R}^{N_p\times N_p}$ corresponding to the kernel framework [20]. As a result, the original objective function can be