Neurocomputing 123 (2014) 121–130

Contents lists available at ScienceDirect

Neurocomputing

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

Incremental min–max projection analysis for classification

Jianwei Zheng, Dan Yang, Shengyong Chen *, Wanliang Wang

College of Computer Science and Technology, Zhejiang University of Technology, No. 288, Liuhe Road, Hangzhou, Zhejiang 310023, China

ARTICLE INFO

Article history:
Received 14 November 2012
Received in revised form
6 April 2013
Accepted 19 June 2013
Communicated by Xiaoqin Zhang
Available online 1 July 2013

Keywords:
Dimensionality reduction
Incremental learning
Singular value decomposition
Orthogonal locally discriminant

ABSTRACT

For data classification, the standard implementation of projection algorithms do not scale well with large dataset size. It makes the computation of large samples infeasible. In this paper, we utilize a block optimization strategy to propose a new locally discriminant projection algorithm termed min–max projection analysis (MMPA). The algorithm takes into account both intra-class and interclass geometries and also possesses the orthogonality property. Furthermore, an incremental MMPA is proposed to learn the local discriminant subspace with newly inserted data by employing the idea of singular value decomposition updating algorithm. Moreover, we extend MMPA to the semi-supervised case and nonlinear case, namely, semi-supervised MMPA and kernel MMPA. The experimental results on image database, hand written digit database, and face database demonstrate the effectiveness of those proposed algorithms.

1. Introduction

The problem of feature extraction is one of the core issues for data mining and classification. Using a more efficient feature extraction method can improve the classification results in the reduced subspace. The problem of dimensionality reduction can be described as follows. Consider a data set $\mathbf{X}$, which consists of n samples $\boldsymbol{x}_i$ ($1 \le i \le n$) in a high-dimensionality space R^m. The objective of dimensionality reduction is to compute a faithful low-dimensionality representation of $\mathbf{X}$, i.e. $\mathbf{Y}=[\boldsymbol{y}_1,\ldots\boldsymbol{y}_n] \in R^{d\times n}$, where $d \ll m$.

Over the past decades, numerous dimension reduction methods have been proposed to find the low-dimensional feature representation. The two most popular techniques for this purpose are principal component analysis (PCA) [1] and linear discriminant analysis (LDA) [2]. PCA is an unsupervised algorithm based on the computation of low-dimensional representation of high dimensional data, which maximizes the total scatter. Comparatively LDA is a supervised feature extraction technique for pattern recognition, and it tends to find a set of projective directions which maximize the between-class scatter and simultaneously minimize the within-class scatter. An intrinsic limitation of LDA is that it usually suffers from the small sample size (SSS) problem, where the sample size is much smaller than the size of dimensionality of samples. Additionally, both PCA and LDA can only see the global Euclidean structure but cannot discover the embedding structure hidden in the high-dimensional data.

In order to exploit the local discriminative manifold structure, a lot of subspace learning techniques have been proposed, such as locality preserving projections (LPP) [3], Maximal Similarity Embedding [4], Local Spline Discriminant Projection [5], and Neighborhood Preserving Embedding [6]. Recently, some researchers pointed out that enforcing an orthogonality relationship between projection directions can achieve competitive effectiveness, and therefore the orthogonal neighborhood preserving projection (ONPP) was introduced [5,7]. However, for classification problems, ONPP and LPP (even in a supervised setting) only focus on the intra-class geometrical information while the interaction of samples from different classes is ignored.

More recently, numerous algorithms have been proposed which take the intra-class preserving into consideration as well as the interclass discriminant [8–11]. Among them, the locality sensitive discriminant analysis (LSDA) [8] and its variation maximum margin projection (MMP) [9] are two typical examples, which gain more competitive results in image recognition applications. Yan et al. [12] explained most of these manifold learning techniques as a general framework that can be defined in a graph-embedding way. Generally, a discriminative feature extraction algorithm is summarized as a graph-based constraint embedding by defining the intrinsic and penalty graphs. In other words, it finds a set of projection directions in the linear embedded subspace, i.e., $J(\mathbf{U})=\mathrm{argmin}\{(\mathbf{U}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{U})/(\mathbf{U}^T\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{U})\}$ or $J(\mathbf{U})=\mathrm{argmin}\{\mathbf{U}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{U}\}$, subject to $\mathbf{U}^T\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{U}=c$, where c is a constant, $\mathbf{X}$ is the data matrix, and $\mathbf{L}$ is the Laplacian matrix of intrinsic graph, which is defined as follows: $\mathbf{L}=\mathbf{D}-\mathbf{W}$, $\mathbf{D}_{ii}=\sum_j \mathbf{W}_{ij}$. Here, $\mathbf{W}$ is the affinity matrix of the intrinsic graph. In addition, $\mathbf{B}$ can be the Laplacian matrix of penalty graph, $\mathbf{B}=\mathbf{D}_p-\mathbf{W}_p$, where $\mathbf{W}_p$ indicates the adjacency matrix of penalty graph. $\mathbf{W}_p$ describes the similarity of interclass data which should be avoided for classification. $\mathbf{D}_p$ is the

* Corresponding author. Tel.: +86 85290309.
E-mail address: sy@ieee.org (S. Chen).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.06.010

diagonal matrix defined in the graph-embedding framework. The general solution of the optimal $\mathbf{U}$ is to find the eigenvector corresponding to the smallest eigenvalue using the generalized eigenvalue decomposition (ED) $\mathbf{XLX}^{\mathrm{T}}\mathbf{U}=\lambda\mathbf{XBX}^{\mathrm{T}}\mathbf{U}$, which has a heavy computation burden because of the high data dimensionality, especially in image and video applications.

Incremental learning has already attracted much attention as a result of the increasing demand for developing machine vision/intelligent systems. Numerous incremental learning algorithms have been proposed, especially in the data-mining domain and the image-retrieval field [13–15]. Most of these recent works are designed for incremental principal component analysis [16,17] and incremental linear discriminant analysis [18,19]. Both of them are global statistic feature extraction algorithms. To our best knowledge, there are few works focusing on the incremental local discriminant embedding except for the ILDSE proposed by Miao et al. [20], which demands that $\mathbf{B}$ must be Laplacian matrix.

In this paper, we propose a new algorithm termed min–max projection analysis (MMPA) based on the perspective of block optimization [21]. MMPA offers three main benefits (1) the algorithm takes into account both intra-class and interclass geometries so that it can achieve better performance in classification; (2) the algorithm produces the orthogonal projection matrix; and (3) the combination matrix of this algorithm can be iteratively computed for the newly inserted samples.

Furthermore, an incremental MMPA is introduced to learn the discriminative sub-manifold structure incrementally, namely, incremental MMPA (IMMPA). This paper also extends MMPA to the semisupervised case and nonlinear case, termed semisupervised MMPA (SMMPA) and kernel MMPA (KMMPA) respectively. SMMPA is produced by incorporating the additional unlabeled samples, and KMMPA performs MMPA in reproducing kernel hilbert space (RKHS). They are powerful. For generalization, the proposed algorithm is also based on graph-embedding framework [12] that incorporates the graph adjacency to represent the discriminative weights of data.

The rest of the paper is organized as follows. Section 2 introduces MMPA algorithm as well as the incremental implementation. Subsequently, the semisupervised MMPA algorithm is proposed in Section 3. In Section 4, the algorithm is extended to the nonlinear case, termed KMMPA. The experimental performance of the proposed algorithms is presented in Section 5. Finally, we conclude this paper in Section 6.

2. Min-max projection analysis (MMPA)

For a given training set $\mathbf{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n]\in R^{m\times n}$, where m, n denote the dimension and the number of the original samples respectively. The proposed MMPA algorithm aims at learning a linear transformation matrix $\mathbf{U}$, which can be used as $\mathbf{Y}=\mathbf{U}^{\mathrm{T}}\mathbf{X}$ to projection the original samples to subspace data $\mathbf{Y}=[\boldsymbol{y}_1,\ldots\boldsymbol{y}_n]\in R^{d\times n}$, where $d\ll m$.

After the transformation, the considered pairwise samples within the same class are as close as possible, while those between classes are as far as possible. The whole algorithm operates in two stages, i.e., block optimization and combination strategy. Block optimization involves building a block using each sample and some of its related samples in the training set. Combination strategy involves combining all the block optimizations to form the final global coordinate using alignment trick [22].

2.1. Block optimization

Suppose that the class label of each sample $\boldsymbol{x}_i$ is denoted by $c(\boldsymbol{x}_i)$. For any given sample $\boldsymbol{x}_i\in\mathbf{X}$, we find its intraclass farthest neighbors outside a certain radius ε_w based on their distance to $\boldsymbol{x}_i$. On the contrary, the interclass nearby points of $\boldsymbol{x}_i$ are searched within the region bounded by the distance ε_b. Assume that the sets of intra-class and interclass neighbors of $\boldsymbol{x}_i$ are indicated by $N_w(\boldsymbol{x}_i)$ and $N_b(\boldsymbol{x}_i)$, respectively. We have

$$N_w(\boldsymbol{x}_i)=\{\boldsymbol{x}_{ij}^w, ||\boldsymbol{x}_{ij}^w-\boldsymbol{x}_i||^2>\varepsilon_w, c(\boldsymbol{x}_{ij}^w)=c(\boldsymbol{x}_i)\}$$

$$N_b(\boldsymbol{x}_i)=\{\boldsymbol{x}_{ij}^b, ||\boldsymbol{x}_{ij}^b-\boldsymbol{x}_i||^2<\varepsilon_b, c(\boldsymbol{x}_{ij}^b)\neq c(\boldsymbol{x}_i)\}$$

where $\boldsymbol{x}_{ij}{}^w$ is one of the farthest neighbors with the same class label as $\boldsymbol{x}_i$ and $\boldsymbol{x}_{ij}{}^b$ is one of the nearest neighbors with different class label as $\boldsymbol{x}_i$. Note that ε_w and ε_b can be different from each other. Then, the intra-class and interclass affinity weight can be defined as follows:

$$w_{ij}^w=\begin{cases}\exp\left(\frac{-||\boldsymbol{x}_i-\boldsymbol{x}_j||^2}{2\sigma^2}\right), & if\ \ x_j\in N_w(\boldsymbol{x}_i)\\ 0, & otherwise\end{cases} \tag{1}$$

$$w_{ij}^b=\begin{cases}\exp\left(\frac{-||\boldsymbol{x}_i-\boldsymbol{x}_j||^2}{2\sigma^2}\right), & if\ \ x_j\in N_b(\boldsymbol{x}_i)\\ 0, & otherwise\end{cases} \tag{2}$$

By combining $\boldsymbol{x}_i$, $N_w(\boldsymbol{x}_i)$ with $N_b(\boldsymbol{x}_i)$, we can build the block for the sample $\boldsymbol{x}_i$ as

$$\mathbf{X}_i=\{\boldsymbol{x}_i\}\cup N_w\cup N_b=\{\boldsymbol{x}_i,\boldsymbol{x}_{i1}^w,\cdots,\boldsymbol{x}_{ik_{iw}}^w,\boldsymbol{x}_{i1}^b,\cdots,\boldsymbol{x}_{ik_{ib}}^b\} \tag{3}$$

where k_{iw}, k_{ib} are the numbers of the farthest within class neighbors and the nearest between class neighbors respectively. The output of each block in the low-dimensional space is denoted by

$$\mathbf{Y}_i=\{\boldsymbol{y}_i,\boldsymbol{y}_{i1}^w,\cdots,\boldsymbol{y}_{ik_{iw}}^w,\boldsymbol{y}_{i1}^b,\cdots,\boldsymbol{y}_{ik_{ib}}^b\} \tag{4}$$

In the subspace, we expect to make the farthest within class neighbors be as close as possible. In the meantime, we expect that the distances between the given sample and the neighbor samples with different labels are as large as possible. Fig. 1 illustrates the process of block optimization, where the solid circle has the radius of ε_w and the outside dotted circle's radius is ε_b. The blue circle on the left represents the ith block in an original high-dimensional space, and the block contains samples with the same label, some of which are the farthest intra-class neighbors (i.e., $\boldsymbol{x}_{i1}{}^w$, $\boldsymbol{x}_{i2}{}^w$, and $\boldsymbol{x}_{i3}{}^w$). The block also contains neighbor samples with different labels (i.e., $\boldsymbol{x}_{i1}{}^b$, $\boldsymbol{x}_{i2}{}^b$, and $\boldsymbol{x}_{i3}{}^b$).The expected results on the block in the low-dimensional space are shown in Fig. 1 (right), where $\boldsymbol{y}_{i1}{}^w$, $\boldsymbol{y}_{i2}{}^w$, and $\boldsymbol{y}_{i3}{}^w$ are as close as possible to $\boldsymbol{y}_i$, whereas $\boldsymbol{y}_{i1}{}^b$, $\boldsymbol{y}_{i2}{}^b$, and $\boldsymbol{y}_{i3}{}^b$ are as far away as possible from $\boldsymbol{y}_i$.

MMPA assumes that the farthest neighbors with the same label $\boldsymbol{y}_{i1}{}^w,\ldots\boldsymbol{y}_{ikiw}{}^w$, are as close as possible to the given sample $\boldsymbol{y}_i$. To make this happen, we minimize the sum of the distances between $\boldsymbol{y}_i$ and $\boldsymbol{y}_{i1}{}^w,\ldots\boldsymbol{y}_{ikiw}{}^w$, and so we have

$$\underset{\boldsymbol{y}_i}{\operatorname{argmin}}\sum_{j=1}^{k_{iw}}||\boldsymbol{y}_i-\boldsymbol{y}_{ij}^w||^2w_{ij}^w \tag{5}$$

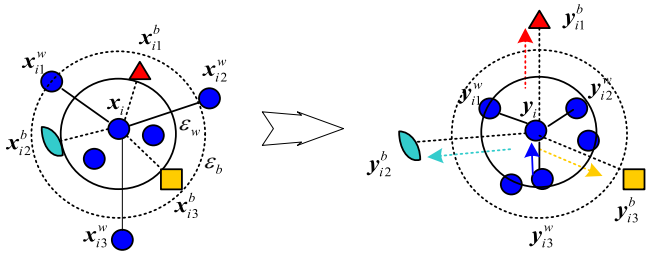

Fig. 1. Block optimization of MMPA.