Neurocomputing 158 (2015) 216–224

Contents lists available at ScienceDirect

Neurocomputing

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

Fast orthogonal linear discriminant analysis with application to image classification

Qiaolin Ye [a,*], Ning Ye [a], Tongming Yin [b]

[a] College of Information Science and Technology, Nanjing Forestry University, Nanjing, PR China
[b] College of Forest Resources and Environment, Nanjing Forestry University, Nanjing, PR China

ARTICLE INFO

Article history:
Received 28 October 2013
Received in revised form
17 November 2014
Accepted 23 January 2015
Communicated by T. Heskes
Available online 3 February 2015

Keywords:
Linear discriminant analysis
Orthogonal linear discriminant analysis
Orthogonal projection vectors
QR decomposition

ABSTRACT

Compared to linear discriminant analysis (LDA), its orthogonalized version is a more effective statistical learning tool for dimension reduction, which devotes to better separating the data points from different classes in the lower-dimensional subspace. However, existing orthogonalized LDA techniques suffer from various drawbacks, including the requirement for expensive computing time. This paper develops an efficient orthogonal dimension reduction approach, referred to as fast orthogonal linear discriminant analysis (FOLDA), which is based on existing orthogonal linear discriminant analysis (OLDA) algorithms. However, different from previous efforts, the new approach applies the QR decomposition and the regression to solve for a new orthogonal projection vector at each iteration, leading to the by far cheaper computational cost. FOLDA achieves comparable recognition rate to existing OLDA algorithms due to the incorporation of the idea and spirit behind the latter ones. Experimental results on image databases, such as MINST, COIL20, MEPG-7 and OUTEX, show the effectiveness and efficiency of our algorithm.

1. Introduction

Techniques for dimension reduction can be implemented to seek for such an optimal low-dimensional space that is helpful for mitigating the so-called "curse of dimensionality" and improving the performance of any classifier, etc. Linear DR finds a meaningful lower-dimensional subspace which provides a compact representation of higher-dimensional data when the data structure is linear in the input space [1,2]. Two most notable linear dimension reduction techniques are principal component analysis (PCA) [1] and linear discriminant analysis (LDA) [3] that have gained wide applications in computer vision and pattern recognition because of their relative simplicity and effectiveness [1,4–6]. Many comparative studies between LDA and PCA on image classification or recognition were made by numerous researchers [3,5–7], in which the results demonstrated that in the terms of recognition rates LDA outperformed PCA significantly [8], implying that it is important for satisfactory design of any classifier to incorporate the supervised information in DR. Thus, LDA can be applied to a family of pattern recognition problems [1,6,9].

The central idea of the classical LDA is to find the optimal projection or transformation that better separates different classes. This optimal projection is obtained by maximizing the between-class dispersion and simultaneously minimizing the within-class dispersion, thus achieving the discrimination between classes. The objective function in the classical LDA is a trace-ratio problem, in which the optimal projection can be computed by a generalized eigenvalue problem. Due to the discrimination of images from different classes, LDA has a direct connection to classification. Despite the effectiveness and applicability, there are many serious limitations in the classical LDA, resulting in many extensions and improvements (we can only cite the most significant ones). The most well-known one is the undersampled or singularity problem in many applications, such face recognition [3], where the dimension of feature space is much larger than the size of training set. Over the past decade, many algorithms have been proposed to solve this problem. In the research [3], Belhumeur et al. proposed to apply LDA after PCA. The authors in [10] used LDA after Singular Value Decomposition (SVD). A common aspect of these two approaches is to perform LDA after another stage of dimension reduction. Since the rank of the within-class scatter matrix $\mathbf{S}_w$ is upper bounded by $m-c$, the maximum dimension of the PCA (or SVD) should be reduced to $m-c$, where m is the size of training set and c denotes the size of classes. However, there is a serious problem in PCA+LDA, which is that the most discriminant information may be lost [11]. To mitigate this problem, there are researchers who suggest keeping the most energy of in the PCA stage [8,12]. Another way to solve the singularity problem in classical LDA is to add the positive

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2015.01.045
0925-2312/

constants to the diagonal elements of $\mathbf{S}_w$ [13]. These algorithms, like classical LDA, transform the trace ratio problems into the ratio trace problems, leading to the non-optimal solution [14].

In [15], Duchene et al. proposed orthogonal linear discriminant analysis (OLDA). OLDA enforces an orthogonality relationship between the discriminant projection vectors to eliminate the redundant information in projection, thus achieving the more powerful discriminant projection vectors than the classical ones in the terms of recognition rates. They adopt a well-designed iterative procedure, and in each iteration, they aim at solving the primal eigenvalue problem of LDA under such an imposed constraint that a new projection vector to be calculated is orthogonal to all the previously obtained projection vectors. Since the projection vectors are independent of the size of classes, there is not such a limitation that the number of projection vectors available is limited to $c-1$. Similar to OLDA, Xiang et al. extended LDA to recursive fisher linear discriminant (RFLD) [8] in which all projection vector are obtained recursively, step by step. Different from OLDA which directly solves a Rayleigh quotient problem with an orthogonality constraint by employing the Lagrange multiplier method, RFLD first rewrites the LDA eigenvalue problem as a generalized eigen-equation, i.e., $\mathbf{S}_b\mathbf{w}_k=\lambda\mathbf{S}_w\mathbf{w}_k$, and then combines the orthogonality constraint with this equation in each iteration, where $\mathbf{S}_w$ and $\mathbf{S}_b$ denote the within-class scatter matrix and the between-class scatter matrix, respectively. Eventually, RFLD still solves a generalized eigen-equation problem in each iteration. It is necessary to note that before new projection vector is computed, the information represented by the previous ones is eliminated from all the samples. Despite the effectiveness of RFLD, RFLD, like OLDA, is expensive computationally, due to that each iteration involves both eigen-decomposition and many operations of matrix inverses as well as matrix multiplications. Still one orthogonal linear discriminant algorithm is maximum margin criterion (MMC) [16], which casts the Rayleigh quotient formulation of the classical LDA as the difference formulation. In addition to establishing the orthogonality relationship between projection vectors, MMC can avoid the singularity problem. Like LDA, it can only extract at most $C-1$ meaningful features [17]. Both OLDA and RFLD permit to define a best discriminant vector, orthogonal to a set of the previously computed vectors, without using any statistical property of this set [15], which is in contrast to MMC. Furthermore, when the dimensionality in the input space is large, it is not infeasible to apply MMC due to the expensive computation resulting from the solution to the formulated large-scale eigenvalue problem.

LDA is to solve the eigen-decomposition problem, which is computationally expensive. To speed up the computation of the LDA problem, Cai et al. proposed spectral regression (SR) [18]. The core idea of SR resort to two separate strategies by first producing response vectors without needing to solve the eigen-decomposition problem and then finding the projection vectors by a regularized least squares formulation which aims to approximate to the response vectors. The computational advantage over LDA is justified by experiments on large image databases. In this paper, we develop a novel algorithm for discriminant analysis, referred to as fast orthogonal linear discriminant analysis (FOLDA), which is essentially based on RFLD [8]. Like RFLD [8], the new approach seeks for the orthogonal projection vectors iteratively. According to some unique properties of matrix, the solution is empirically obtained without the need to solve the eigenvalue problem. Then, the spectral regression [18] is used to obtain a new orthogonal projection vector. Clearly, the process of solution to the FOLDA problem does not involve the eigen-decomposition, multiple matrix inverses, and multiplications, leading to the less computational cost than RFLD. FOLDA does not use any statistical property of the previously obtained orthogonal projection vectors and is permitted to define a "best discriminant" vector, orthogonal to these orthogonal projection vectors. Therefore, there is no limitation on the number of projection vectors available from FOLDA, which is in contrast to MMC. We also demonstrate the efficiency of FOLDA by analyzing and comparing the time complexities of existing orthogonal methods. The experiment is tried out on four image databases, such as MEPG-7, COIL20, MIST, and OUTEX indicates the effectiveness and efficiency of our proposed FOLDA algorithm.

2. Related works

In this section, we briefly review LDA and its two orthogonal extensions, such as RFLD [8] and MMC [16]. We denote the sample set as $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n]$, $\mathbf{x}_i\in\mathbb{R}^d$, where n is the sample size and d the feature dimensionality. The class label of the sample $\mathbf{x}_i$ is from the set $\{1,2,\ldots,c\}$, where c is the number of classes. Define n_l as the number of the labeled samples from the lth class. Let $\mathbf{W}^{(k-1)}=(\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_{k-1})\in\mathrm{R}^{d\times(k-1)}$ be a set of previously-computed $k-1$ orthogonal projection basis vectors. Define by $\mathbf{z}\in\mathbb{R}^r(1\le r\le d)$ a low-dimensional representation of a high-dimensional sample $\mathbf{x}$ in the original input space, where r is the dimensionality of the reduced space. The purpose of DR is to seek for a transformation matrix $\mathbf{W}$, such that a lower representation $\mathbf{z}$ of the sample $\mathbf{x}$ can be yielded as $\mathbf{z}=\mathbf{W}^T\mathbf{x}$, where T denotes the transpose.

2.1. Linear discriminant analysis (LDA)

LDA seeks for projection vectors on which the data points from different classes are well separated. The objective function of LDA is as follows:

$$\mathbf{W}^*=\arg\ \max\frac{tr(\mathbf{W}^T\mathbf{S}_b\mathbf{W})}{tr(\mathbf{W}^T\mathbf{S}_w\mathbf{W})} \tag{1}$$

or

$$\mathbf{W}^*=\arg\ \max\frac{tr(\mathbf{W}^T\mathbf{S}_b\mathbf{W})}{tr(\mathbf{W}^T\mathbf{S}_t\mathbf{W})} \tag{2}$$

where $\mathbf{S}_w=\sum_{l=1}^{c}\sum_{i=1}^{n_l}(\mathbf{x}_i-\boldsymbol{\mu}_l)(\mathbf{x}_i-\boldsymbol{\mu}_l)^T$, $\mathbf{S}_b=\sum_{l=1}^{c}n_l(\boldsymbol{\mu}_l-\boldsymbol{\mu})(\boldsymbol{\mu}_l-\boldsymbol{\mu})^T$, and $\mathbf{S}_t=\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T$ denote the within-class scatter matrix, the between-class scatter matrix, and the global scatter matrix, respectively, in which $\boldsymbol{\mu}_l$ and $\boldsymbol{\mu}$ denote the respective mean of the sample set in lth class and the complete training set. The projection vectors are selected as the eigenvectors corresponding to the first r largest eigenvalues of $(\mathbf{S}_w)^{-1}\mathbf{S}_b$ or $(\mathbf{S}_t)^{-1}\mathbf{S}_b$.

From a graph-embedding viewpoint, the objective function of LDA in (2) is equivalent to [18]

$$\mathbf{W}^*=\arg\ \max\frac{tr(\mathbf{W}^T\overline{\mathbf{X}}\mathbf{V}\overline{\mathbf{X}}^T\mathbf{W})}{tr(\mathbf{W}^T\overline{\mathbf{X}}\overline{\mathbf{X}}^T\mathbf{W})} \tag{3}$$

in which $\mathbf{V}=\begin{pmatrix}\mathbf{V}^{(1)} & 0 & \ldots & 0\\ 0 & \mathbf{V}^{(2)} & \ldots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \ldots & \mathbf{V}^{(c)}\end{pmatrix}$ is a matrix of dimensions $n\times n$, and $\overline{\mathbf{X}}$ denotes the centered data matrix. Here, $\mathbf{V}^{(l)}$, $l=1,2,\ldots,c$ is a $n_l\times n_l$ matrix with the entries equal to $1/n_l$. LDA is non-orthogonal and has a serious limitation, that is, the number of projection vectors available is limited to $c-1$, which limits its applications to a large class of problems [8].