Pattern Recognition 47 (2014) 1698–1708

Contents lists available at ScienceDirect

Pattern Recognition

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

A Rayleigh–Ritz style method for large-scale discriminant analysis

Lin Zhu *, De-Shuang Huang

School of Electronics and Information Engineering, Tongji University, 4800 Caoan Road, Shanghai 201804, China

ARTICLE INFO

Article history:
Received 5 September 2012
Received in revised form
6 September 2013
Accepted 5 October 2013
Available online 20 October 2013

Keywords:
Dimensionality reduction
Linear Discriminant Analysis (LDA)
Generalized Eigenvalue Decomposition (GEVD)
Rayleigh–Ritz procedure

ABSTRACT

Linear Discriminant Analysis (LDA) is one of the most popular approaches for supervised feature extraction and dimension reduction. However, the computation of LDA involves dense matrices eigendecomposition, which is time-consuming for large-scale problems. In this paper, we present a novel algorithm called Rayleigh–Ritz Discriminant Analysis (RRDA) for efficiently solving LDA. While much of the prior research focus on transforming the generalized eigenvalue problem into a least squares formulation, our method is instead based on the well-established Rayleigh–Ritz framework for general eigenvalue problems and seeks to directly solve the generalized eigenvalue problem of LDA. By exploiting the structures in LDA problems, we are able to design customized and highly efficient subspace expansion and extraction strategy for the Rayleigh–Ritz procedure. To reduce the storage requirement and computational complexity of RRDA for high dimensional, low sample size data, we also establish an equivalent reduced model of RRDA. Practical implementations and the convergence result of our method are also discussed. Our experimental results on several real world data sets indicate the performance of the proposed algorithm.

1. Introduction

With the advancements in data collection and storage technologies, there has been an exponential increase in the availability and usage of large, high-dimensional datasets. These data can directly be represented as vectors in high-dimensional vector spaces. Obviously, operating directly on such high-dimensional vector space is ineffective and may lead to high computational and storage demands as well as poor performance. A typical way to circumvent the "curse of dimensionality" problem [1] and other undesired properties of high dimensional spaces is to use dimensionality reduction techniques. The goal of dimensionality reduction is to map a set of high-dimensional samples into a lower dimensional space while preserving the intrinsic structure in the data.

Up to now, researchers have developed a variety of dimensionality reduction methods. Based on how to utilize the label information, these algorithms can be broadly divided into three classes, i.e., unsupervised, supervised and semi-supervised. The first class is supervised which typically includes linear discriminant analysis (LDA) [2,3], maximization of the geometric mean of all divergences (MGMD) [4], max–min distance analysis(MMDA) [5], etc. The second class is unsupervised which include principal component analysis (PCA) [6] and locality preserving projections (LPP) [7,8], etc. The third class is semi-supervised [9,10] which can use unlabeled data for promoting supervised methods. Beyond the commonness in mathematical formulation [11] shared by these algorithms, Zhang et al. proved that most popular dimensionality reduction algorithms, unsupervised or supervised, can be explained as instances of a ubiquitously two-stage framework named "patch alignment" [12].

One of the most popular supervised dimensionality reduction algorithms is linear discriminant analysis (LDA). It has been widely used in many applications such as microarray data classification; face recognition and gait recognition, etc [2,13–15].

Computationally, LDA amounts to solving a generalized symmetric semi-definite eigenvalue decomposition (GEVD) problem. A straightforward implementation can thus be very time-consuming for large datasets [16]. So far a number of methods [3,16–22] have been explored to solve LDA with improved scalability. The key idea of these approaches is to transform the generalized eigenvalue problem into a least squares formulation, which can be solved efficiently using existing algorithms such as LSQR [21,23]. Two representative algorithms are least squares linear discriminant analysis (LS-LDA) [18,19] and spectral regression discriminant analysis (SRDA) [16]. Recently, Sun introduced 2SrLDA [21], a generalization of LS-LDA which relaxes the equivalence conditions and incorporates a regularization term.

Alternatively, one can use general-purpose GEVD algorithms to solve LDA. It is known that for solving small to medium size GEVD

* Corresponding author. Tel.: +86 551 3493367.
E-mail addresses: zhomlynn@ustc.edu.cn, lizhonyx@163.com (L. Zhu).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.10.007

problems, 'Rayleigh–Ritz' methods like Lanczos method and JDQZ are often the ideal choice [2,13–15]. In [21], Sun attempted to solve LDA based on Lanczos method. However, the experiments comparisons show that it is much slower than regression-based methods. This is not surprising because a standard GEVD solver would encounter severe computational bottleneck or even become impractical for large problems. For example, Rayleigh–Ritz type GEVD algorithms generally cannot guarantee finding the largest eigenpairs (which is required by LDA) via short-term recurrences, therefore more sophisticated strategy like trust-reign method has to be employed [24], which can make the computation inefficient. In addition, for symmetric eigenvalue problems, Rayleigh–Ritz methods like Lanczos algorithm can guarantee that the approximate solution at each iteration be computed efficiently using divide-and-conquer method [25]. However this advantage also doesn't exist for general GEVD methods.

In this paper, we propose a novel Rayleigh–Ritz algorithm for LDA called Rayleigh–Ritz discriminant analysis (RRDA). Like JDQZ and Lanczos method, we adopt the 'Rayleigh–Ritz' procedure as a general framework. However, the subspace expansion and extraction strategy is specially designed to gradually optimize the ratio trace objective function of LDA. By exploiting the special structure of LDA problem, the above-mentioned issues of general purpose GEVD methods can also be solved. We demonstrate both theoretically and via numerical examples that this hybrid algorithm efficiently produces a solution of high precision.

Additionally, in many applications, data are highly dimensional, while the number of available training samples is usually much smaller. Such is the case for the image databases of facial recognition, gene expression data, as well as the text documents [26,27]. To reduce the storage requirement and computational complexity of the iteration for these high dimensional, low sample size data, we also establish an equivalent reduced model of RRDA which is of order n instead of order d, where d is the number of features and n is the number of samples. Practical implementations and the convergence result of RRDA are also discussed. Experiments show that RRDA is more efficient than 2SrLDA and SRDA, which are the state-of-the-art methods for solving LDA.

The rest of this paper is organized as follows. Section 2 outlines LDA and the Rayleigh–Ritz procedure. The proposed method is discussed in Section 3. Experimental results on several real world data sets are reported in Section 4. Finally, some concluding remarks are drawn in Section 5.

2. Background

In preparation for our description of the proposed method, we introduce some notations. Throughout this paper all matrices are boldface uppercase, and vectors are boldface lowercase. n is the number of samples, d is the data dimensionality, and c is the number of classes (or labels). The ith sample is denoted as $\boldsymbol{x}_i \in \mathbb{R}^{d\times 1}$, and $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]\in\mathbb{R}^{d\times n}$ represents the data matrix. Without loss of generality, we assume that $\boldsymbol{X}$ is partitioned into c classes as $\boldsymbol{X}=[\boldsymbol{X}_1, \boldsymbol{X}_2,\cdots, \boldsymbol{X}_c]$, where $\boldsymbol{X}_i\in\mathbb{R}^{d\times n_i}$ corresponds to the data points from the ith class. $\boldsymbol{I}_q$ is the q-by-q identity matrix and $\boldsymbol{e}_q$ is a vector of all ones with length q. The centering matrix is defined as $\boldsymbol{C}_q=\boldsymbol{I}_q-\boldsymbol{e}_q\boldsymbol{e}_q^T$. For a matrix $\boldsymbol{A}$, we let $\boldsymbol{A}^+$ be the Moore–Penrose inverse of $\boldsymbol{A}$, trace($\boldsymbol{A}$) be the trace of $\boldsymbol{A}$, rank($\boldsymbol{A}$) be the rank of $\boldsymbol{A}$ and span($\boldsymbol{A}$) be the range space of $\boldsymbol{A}$.

2.1. Linear discriminant analysis

The solution of LDA optimizes the ratio trace problem [21,28,29]:

$$\arg\max_{\boldsymbol{W}^T\boldsymbol{S}_t\boldsymbol{W}=\boldsymbol{I}} \operatorname{trace}[(\boldsymbol{W}^T\boldsymbol{S}_t\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W}] \tag{1}$$

where the total covariance matrix $\boldsymbol{S}_t$ and the between-class covariance matrix $\boldsymbol{S}_b$ are defined as

$$\boldsymbol{S}_t=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T$$
$$\boldsymbol{S}_b=\frac{1}{n}\sum_{k=1}^{c}n_k(\boldsymbol{\mu}_k-\boldsymbol{\mu})(\boldsymbol{\mu}_k-\boldsymbol{\mu})^T \tag{2}$$

where $\boldsymbol{\mu}_i=(1/n_i)\boldsymbol{X}_i\boldsymbol{e}_{n_i}$ denotes the centroid of the class i and $\boldsymbol{\mu}=(1/n)\boldsymbol{X}\boldsymbol{e}_n$ denotes the global centroid.

The optimization problem (1) is solved by computing all the generalized eigenpairs [16]:

$$\lambda_i\boldsymbol{S}_t\boldsymbol{w}_i=\boldsymbol{S}_b\boldsymbol{w}_i,\ \boldsymbol{w}_i\neq 0,\ \lambda_i>0 \tag{3}$$

and thus, the optimal $\boldsymbol{W}$ consists of eigenvectors of $\boldsymbol{S}_t^{-1}\boldsymbol{S}_b$ corresponding to all positive eigenvalues, provided that $\boldsymbol{S}_t$ is nonsingular.

In practical application, to deal with the singularity of $\boldsymbol{S}_t$, many generalizations of LDA have been proposed [15,17,27,28,30–39]. One widely used approach is called Regularized Discriminant Analysis (RDA) [33], which penalizes the total scatter matrix $\boldsymbol{S}_t$ with a regularizer $\alpha\boldsymbol{I}_d$, $\alpha>0$, and the solution is given by solving the GEVD problem

$$\lambda_i(\boldsymbol{S}_t+\alpha\boldsymbol{I}_d)\boldsymbol{w}_i=\boldsymbol{S}_b\boldsymbol{w}_i,\boldsymbol{w}_i\neq 0,\quad \lambda_i>0 \tag{4}$$

Since LDA is a special case of RDA with $\alpha=0$, we mainly discuss RDA in the sequel.

2.2. Rayleigh–Ritz method for GEVD problems

The Rayleigh–Ritz procedure serves as the basic framework of many algorithms for solving linear and nonlinear eigenvalue problems. Notable examples include the Lanczos method [40], JDQZ (including JD-related variants) [41–43] and GY [44].

In order to find the generalized eigenpairs of a given matrix pair ($\boldsymbol{S}_1$, $\boldsymbol{S}_2$), the Rayleigh–Ritz procedure selects the approximate eigenvector from a search subspace span($\boldsymbol{V}_k$) that is expanded iteratively. Each iteration consists of two parts. In the first part, the 'extraction' part, the projected generalized eigenproblem

$$\boldsymbol{V}_k^T\boldsymbol{S}_1\boldsymbol{V}_k\boldsymbol{\eta}=\lambda\boldsymbol{V}_k^T\boldsymbol{S}_2\boldsymbol{V}_k\boldsymbol{\eta} \tag{5}$$

is solved and a solution (λ,$\boldsymbol{\eta}$) is selected. The Ritz value λ and Ritz vector $\boldsymbol{V}_k\boldsymbol{\eta}$ form an approximate generalized eigenvalue and generalized eigenvector, respectively. In the second part, the 'expansion' part, the search space span($\boldsymbol{V}_k$) is enlarged by adding a new basis matrix $\boldsymbol{P}_k$ to it. This process is summarized in Algorithm 1. The idea is that, as the search subspace grows, the eigenpair approximations will converge to an eigenpair of the original problem. In order to keep computation costs low, we usually do not expand the search space to the whole space.

Algorithm 1. The Rayleigh–Ritz framework for solving GEVD Problems

1) **Start:** Matrix pair ($\boldsymbol{S}_1$, $\boldsymbol{S}_2$), choose an initial non-zero matrix $\boldsymbol{V}_1$.
2) **Output:** $\boldsymbol{W}_{opt}$ as the eigenvectors of ($\boldsymbol{S}_1$, $\boldsymbol{S}_2$) with positive eigenvalues.
3) **Iterate:** Until convergence, for $k=1,2,\ldots$ do:
   3.1) **Expansion:** Choose a subspace basis $\boldsymbol{P}_k$ based on certain criteria, expand $\boldsymbol{V}_{k-1}$ with $\boldsymbol{P}_k$ to $\boldsymbol{V}_k$ such that $\boldsymbol{V}_k=[\boldsymbol{V}_{k-1}\ \boldsymbol{P}_k]$.
   3.2) **Extraction:** Compute the positive eigenpairs (λ_i, $\boldsymbol{\eta}_i$) of ($\boldsymbol{V}_k^T\boldsymbol{S}_1\boldsymbol{V}_k$, $\boldsymbol{V}_k^T\boldsymbol{S}_2\boldsymbol{V}_k$) and the Ritz vectors $\boldsymbol{w}_i=\boldsymbol{V}_k\boldsymbol{\eta}_i$, $i=1,2,\cdots,c$, Let $\boldsymbol{W}_k=[\boldsymbol{w}_1\boldsymbol{w}_2\cdots]$;
   3.3) Test for convergence. Stop if satisfied.
4) **End:** Output $\boldsymbol{W}_k$ as the approximation of $\boldsymbol{W}_{opt}$.