Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT

PERGAMON

Pattern Recognition 37 (2004) 851–854

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Rapid and Brief Communication

LDA/QR: an efficient and effective dimension reduction algorithm and its theoretical foundation

Jieping Ye[a,*], Qi Li[b]

[a] *Department of Computer Science and Engineering, University of Minnesota, 200 Union Street, S.E. Minneapolis, MN 55455, USA*
[b] *Department of Computer Science, University of Delaware, USA*

Received 4 August 2003; accepted 12 August 2003

Abstract

LDA/QR, a linear discriminant analysis (LDA) based dimension reduction algorithm is presented. It achieves the efficiency by introducing a QR decomposition on a small-size matrix, while keeping competitive classification accuracy. Its theoretical foundation is also presented.

Keywords: Linear discriminant analysis; QR-decomposition; Pseudo-inverse

1. Introduction

With the efforts of addressing the singularity problem (of scatter matrix) in the classical linear discriminant analysis (LDA), LDA is receiving more and more attentions. PCA+LDA [1] is a popular way to deal with the singularity problem. Recently, LDA/GSVD was developed in Ref. [2] along this direction. Pseudo-inverse [3] is a common way to deal with the singularity problem on matrix. Generalized LDA based on pseudo-inverse was presented in Ref. [4] to overcome the singularity problem in classical LDA.

In this paper, we will present a novel member of LDA family, namely LDA/QR, as an efficient and effective dimension reduction algorithm (by effective we mean the high classification accuracy). The utilization of QR-decomposition on the small-size matrix is the soul of LDA/QR algorithm. It can be shown that its time complexity is linear on the size of the data and also linear on the number of dimensions. It is also numerically stable since all of the decompositions and the inversions are applied to small dimension matrices.

We will also give the theoretical foundation of LDA/QR by showing the equivalence between LDA/QR and the generalized LDA. The experiments in the final part of this paper will show the effectiveness of the LDA/QR dimension reduction algorithm.

2. Classical LDA and generalized LDA

Throughout the paper, N denotes the number of points, n is the dimension, and k is the number of classes. Matrix $A \in R^{n\times N}$ is the data matrix, where each column of A denotes a training data point in the n-dimensional space. $A_i \in R^{n\times N_i}$ is the data matrix containing the data points from the ith class, where N_i is the size of the ith class.

Classical LDA is found by solving one of the trace optimizations in Eq. (1),

$$G = \arg\max_G \operatorname{trace}((G^{\mathrm{T}}S_wG)^{-1}(G^{\mathrm{T}}S_bG)) \quad \text{if } S_w \text{ is nonsingular}$$

or

$$G = \arg\min_G \operatorname{trace}((G^{\mathrm{T}}S_bG)^{-1}(G^{\mathrm{T}}S_wG)) \quad \text{if } S_b \text{ is nonsingular.} \tag{1}$$

* Corresponding author. Tel.: +1-612-626-7504; fax: +1-612-625-0572.
E-mail addresses: jieping@cs.umn.edu (J. Ye), qili@mail.eecis.udel.edu (Q. Li).

0031-3203/$30.00 © 2003 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2003.08.006

The between- and within-class matrices S_b and S_w in Eq. (1) are defined as [4]

$$S_b = \frac{1}{N}\sum_{i=1}^{k} N_i(m_i - m)(m_i - m)^{\mathrm{T}} = \frac{1}{N} H_b H_b^{\mathrm{T}},$$

$$S_w = \frac{1}{N}\sum_{i=1}^{k}\sum_{x\in C_i}(x - m_i)(x - m_i)^{\mathrm{T}} = \frac{1}{N} H_w H_w^{\mathrm{T}}, \qquad (2)$$

where

$$H_b = \left[\sqrt{N_1}(m_1 - m), \ldots, \sqrt{N_k}(m_k - m)\right] \in R^{n\times k},$$
$$H_w = [A_1 - m_1 \cdot e_1, \ldots, A_k - m_k \cdot e_k] \in R^{n\times N}, \qquad (3)$$

$e_i = (1, \ldots, 1) \in R^{1\times N_i}$, A_i is the data matrix containing only the points from the ith class, m_i is the centroid of class C_i, and m is the global centroid of the whole data set. Note the matrix H_b is much smaller than H_w.

The optimization problem in Eq. (1) is equivalent to the following generalized eigenvalue problem, $S_b x = \lambda S_w x$, for $\lambda \neq 0$. The solution can be obtained by solving an eigenproblem on matrix $S_w^{-1}S_b$, if S_w is nonsingular, or on $S_b^{-1}S_w$, if S_b is nonsingular. There are atmost $k-1$ eigenvectors corresponding to nonzero eigenvalues, since the rank of the matrix S_b is bounded above by $k-1$. Therefore, the reduced dimension by classical LDA is atmost $k-1$.

Generalized LDA uses pseudo-inverse to deal with singularity problem. A natural extension of classical LDA, using the pseudo-inverse, is to solve the eigenproblem on $S_b^{+}S_w$ or $S_w^{+}S_b$. The pseudo-inverse of a matrix can be computed by SVD [3]. More specifically, if $A = U\Sigma V^{\mathrm{T}}$ is the singular value decomposition of the matrix $A \in R^{m\times n}$, where $U \in R^{m\times t}$ and $V \in R^{n\times t}$ have orthonormal columns, $\Sigma \in R^{t\times t}$ is diagonal with positive diagonal entries, and $t = \mathrm{rank}(A)$, then the pseudo-inverse of A is defined as $A^{+} = V\Sigma^{-1}U^{\mathrm{T}}$.

3. LDA/QR algorithm for dimension reduction

In this section, we present the LDA/QR algorithm. It has two stages. The essence of the first stage is the maximum separability among different classes obtained by QR-decomposition [3] (note that QR is more efficient than SVD numerically). The second stage contains four steps that involve the concern of within-class distance.

The first stage aims to solve the following optimization problem:

$$G = \arg\max_{G^{\mathrm{T}}G=I} \mathrm{trace}(G^{\mathrm{T}}S_b G). \qquad (4)$$

Note that this optimization problem only gives the concern on maximizing between-class distance. The solution can be obtained by solving the eigenproblem on S_b. However, the solution to Eq. (4) can also be obtained through QR-decomposition on the matrix H_b as follows. Let $H_b = QR$ be the QR-decomposition on H_b, where $Q \in R^{N\times t}$ has orthonormal columns, $R \in \mathbb{R}^{t\times k}$ is an upper triangular matrix, and $t = \mathrm{rank}(H_b)$, then $G = QW$, for any orthogonal matrix $W \in \mathbb{R}^{t\times t}$ solves the optimization problem in Eq. (4). Detailed proof is omitted because of space limit.

Note the rank t of the matrix H_b, is bounded above by $k-1$. In practice, the k centroids in the data set are usually linearly independent. In this case, the reduced dimension t equals to $k-1$.

The LDA/QR algorithm concerns the within-class distance at the second stage. Its second stage in-cooperates the within-class scatter information by applying a relaxation scheme on W (relaxing W from an orthogonal matrix to be an arbitrary matrix, details following in the next paragraph). The final optimization problem is exactly the same one as in classical LDA, but with matrices of much smaller size, hence can be solved efficiently and stably.

Specifically, we make a relaxation on the solution to the optimization problem (4). That is, we look for a transformation matrix G such that $G = QW$, for any matrix $W \in \mathbb{R}^{t\times t}$, hence W is not required to be orthogonal. The original problem of finding G is now equivalent to computing W. The first stage of the LDA/QR algorithm choose W to be any orthogonal matrix, by omitting the within-class scatter. The second stage of LDA/QR algorithm solves this limitation by considering both the between-class and within-class scatters as follows. Since

$$G^{\mathrm{T}}S_b G = W^{\mathrm{T}}(Q^{\mathrm{T}}S_b Q)W, \quad G^{\mathrm{T}}S_w G = W^{\mathrm{T}}(Q^{\mathrm{T}}S_w Q)W, \qquad (5)$$

the original optimization on finding optimal G is equivalent to finding W, such that

$$W = \arg\min_{W} \mathrm{trace}((W^{\mathrm{T}}\tilde{S}_b W)^{-1}(W^{\mathrm{T}}\tilde{S}_w W)), \qquad (6)$$

where $\tilde{S}_b = Q^{\mathrm{T}}S_b Q$ and $\tilde{S}_w = Q^{\mathrm{T}}S_w Q$. Note that $\tilde{S}_b$ is nonsingular, hence the optimization problem in Eq. (6) is well defined and can be solved using similar method for the optimization problem in Eq. (1) of classical LDA. That is, we compute optimal W, by solving a small eigenproblem on $\tilde{S}_b^{-1}\tilde{S}_w$. The main steps of LDA/QR include: (1) construct the matrix H_b and H_w as in Eq. (3); (2) compute QR-decomposition of H_b by $H_b = QR$; (3) compute the t eigenvectors w_i of $\tilde{S}_b^{-1}\tilde{S}_w$, with increasing eigenvalues, where $\tilde{S}_b$ and $\tilde{S}_w$ are defined above; and (4) $G \leftarrow QW$, where $W = [w_1, \ldots, w_t]$. It is easy to check that the time complexity of LDA/QR is $O(Nnk)$.

4. Equivalence between LDA/QR and generalized LDA

As discussed in Section 2, classical LDA computes the optimal transformation matrix by solving the eigenproblem on $S_w^{-1}S_b$, if S_w is nonsingular, or $S_b^{-1}S_w$, if S_b is nonsingular. A natural extension of the classical LDA, using the pseudo-inverse, is to solve the eigenproblem on $S_b^{+}S_w$ or $S_w^{+}S_b$, as discussed in Section 2 on generalized LDA.

Interestingly, the solution to the eigenproblem on $S_b^{+}S_w$ is equivalent to the solution by the LDA/QR algorithm, as stated in the following theorem.

Theorem 4.1. *Let G be the optimal transformation matrix obtained from LDA/QR algorithm. Then the columns of G are eigenvectors of $S_b^+ S_w$ corresponding to nonzero eigenvalues.*

Proof. Let x be an eigenvector of $S_b^+ S_w$ corresponding to nonzero eigenvalue λ, i.e. $S_b^+ S_w x = \lambda x$. Let $H_b = [Q, \tilde{Q}] \begin{pmatrix} R \\ 0 \end{pmatrix}$ be the QR-decomposition of H_b, where $[Q, \tilde{Q}] \in R^{n \times n}$ is orthogonal, $Q \in R^{n \times t}$, $\tilde{Q} \in R^{n \times n-t}$, $R \in R^{t \times k}$ is upper triangular and $t = \mathrm{rank}(H_b)$. It follows that

$$S_b^+ = (H_b H_b^{\mathrm{T}})^+ = \left([Q, \tilde{Q}] \begin{pmatrix} RR^{\mathrm{T}} & 0 \\ 0 & 0 \end{pmatrix} [Q, \tilde{Q}]^{\mathrm{T}} \right)^+$$

$$= [Q, \tilde{Q}] \begin{pmatrix} (RR^{\mathrm{T}})^{-1} & 0 \\ 0 & 0 \end{pmatrix} [Q, \tilde{Q}]^{\mathrm{T}}.$$

Hence

$$S_b^+ S_w x = [Q, \tilde{Q}] \begin{pmatrix} (RR^{\mathrm{T}})^{-1} & 0 \\ 0 & 0 \end{pmatrix} [Q, \tilde{Q}]^{\mathrm{T}} H_w H_w^{\mathrm{T}} x = \lambda x.$$

It follows that

$$\begin{pmatrix} (RR^{\mathrm{T}})^{-1} \\ 0 \end{pmatrix} Q^{\mathrm{T}} H_w H_w^{\mathrm{T}} [Q, \tilde{Q}] \begin{pmatrix} Q^{\mathrm{T}} \\ \tilde{Q}^{\mathrm{T}} \end{pmatrix} x = \lambda \begin{pmatrix} Q^{\mathrm{T}} \\ \tilde{Q}^{\mathrm{T}} \end{pmatrix} x.$$

It is easy to check $\tilde{Q}^{\mathrm{T}} x = 0$. Hence $(RR^{\mathrm{T}})^{-1}(Q^{\mathrm{T}} H_w H_w^{\mathrm{T}} Q) Q^{\mathrm{T}} x = \lambda Q^{\mathrm{T}} x$, which implies $Q^{\mathrm{T}} x$ is an eigenvector of $(RR^{\mathrm{T}})^{-1} Q^{\mathrm{T}} H_w H_w^{\mathrm{T}} Q$, the same matrix used in step 3 of LDA/QR algorithm. This completes the proof of this theorem. □

5. Experiments and discussion

We presented the experimental results on two kinds of data sets. The first category is of image data, including AR and ORL. The second category is of text data, derived from the TREC collections, and the *Reuters-21578* text categorization test collection Distribution 1.0. We considered the first stage of the LDA/QR algorithm as a separate dimension algorithm and named it pre-LDA/QR. We did comparison between PCA+LDA, pre-LDA/QR, LDA/QR and LDA/GSVD.

The K-nearest-neighbor (KNN) algorithm was applied to evaluate the quality of different dimension-reduction algorithms as in Refs. [2,4]. The classification accuracies are estimated by 10-fold cross validation. The accuracy curves associated with different algorithms on each data set is presented in Fig. 1 whose horizontal axis shows the size of neighbors used in KNN and whose vertical axis shows the accuracy.

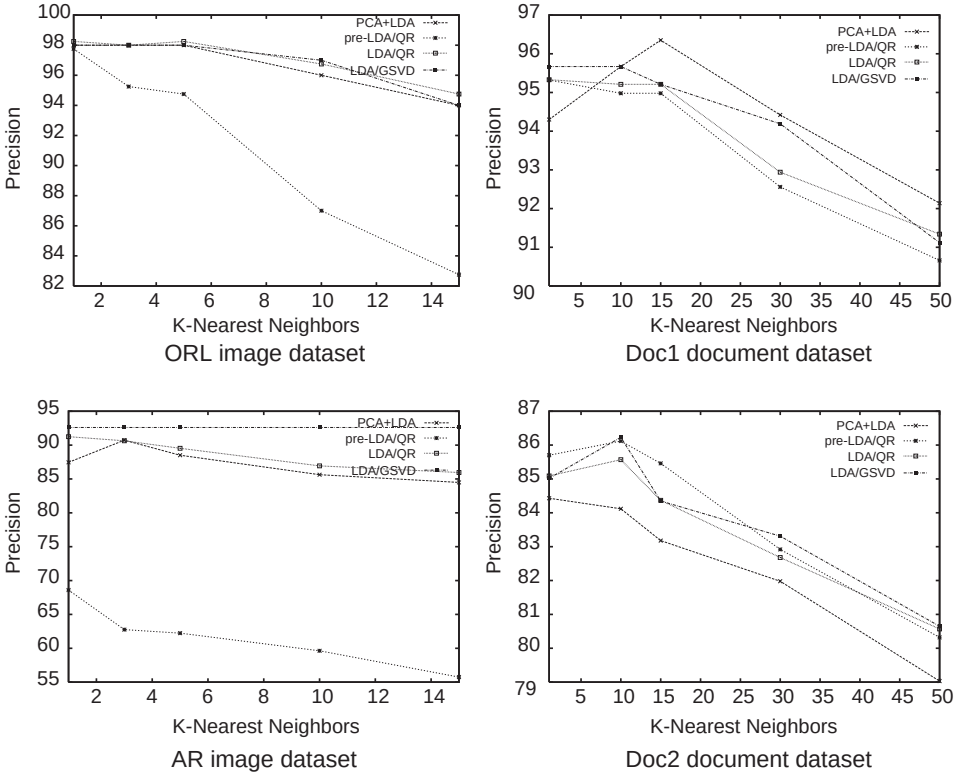

Fig. 1. The x-axis is the number of neighbors used in KNN, and the y-axis is the accuracy.

ORL face data set, is a well-known data set for face recognition. It contains the face images of 40 persons. The image size is 92×112. The AR face image data, is not only huge, but also pretty large area of occlusion, such as sun glasses and scarf. The existence of occlusion dramatically increases the within-class variances of AR face image data. We use a subset of AR face. This subset contains 1638 face images of 126 people's faces. Its image size is 768×576. We first crop the image from the row 100 to 500, and column 200 to 550, and then subsample the cropped images with sample step 4×4. The dimension of each instance is thus 8888.

The Doc1 document data set is a text document data set derived from the TREC collections (http://trec.nist.gov). It has 878 documents belonging to 10 classes, with 7455 dimensions. The Doc2 document data set is a text document data set derived from *Reuters-21578* text categorization test collection Distribution 1.0 (http://www.research.att.com/~lewis). It has 1657 documents belonging to 24 classes, with 3759 dimensions. We used the *tf-idf* weighting scheme [4] for the text data.

The most interesting result lies in the AR image data set. We can observe that LDA/QR and LDA/GSVD distinctly outperform the other dimension reduction algorithms. Remind that AR face images contain pretty large area of occlusion whose direct effect is the large within-class variances of each class/individual. The effort of minimizing of the within-class variance achieves distinct success in this situation. Pre-LDA/QR does not have the effort in minimizing the within-class variance. For the two-stage PCA+LDA algorithm, we use the first 150 principal components [1] (note the reduced dimension in this case is $126 - 1$). On ORL, the best accuracies are around 99%. Multiple LDA members can achieve this accuracy. This is mainly due to the relatively small within-class variances in this data.

The main observation on the text data is that pre-LDA/QR, a member in LDA without any effort in handling the within-class variance, performs pretty well especially when more than 15 neighbors. This observation reminds us a fact on text data, i.e., they do have relatively small within-class variance.

References

[1] D.L. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Pattern Anal. Mach. Intell. 18 (8) (1996) 831–836.

[2] P. Howland, M. Jeon, H. Park, Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition, SIAM J. Matrix Anal. Appl. 25 (1) (2003) 165–179.

[3] G.H. Golub, C.F.V. Loan, Matrix Computations, 3rd Edition, The Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[4] J. Ye, R. Janardan, C. Park, H. Park, A new optimization criterion for generalized discriminant analysis on undersampled problems, Technical Report TR03-026, Department of Computer Science, University of Minnesota, 2003.

[1] We use the first 100 principal components for the other data sets.