A Direct LDA Algorithm for High-Dimensional Data – with Application to Face Recognition

Hua Yu [1], Jie Yang

Interactive System Labs, Carnegie Mellon University, Pittsburgh, PA 15213

1. Introduction

Linear Discriminant Analysis (LDA) has been successfully used as a dimensionality reduction technique to many classification problems, such as speech recognition, face recognition, and multimedia information retreival. The objective is to find a projection A that maximizes the ratio of between-class scatter S_b against within-class scatter S_w (Fisher's criterion):

$$\arg\max_A \frac{|AS_bA^T|}{|AS_wA^T|}$$

However, for a task with very high dimensional data such as images, the traditional LDA algorithm encounters several difficulties. Consider face recognition for example. A low-definition face image of size 64 by 64 implies a feature space of $64 \times 64 = 4096$ dimensions, and therefore scatter matrices of size $4096 \times 4096 = 16$M. First, it is computationally challenging to handle big matrices (such as computing eigenvalues). Second, those matrices are almost always singular, as the number of training images needs to be at least 16M for them to be non-degenerate.

Due to these difficulties, it is commonly believed that a direct LDA solution for such high-dimensional data is infeasible. Thus, ironically, before LDA can be used to reduce dimensionality, another procedure has to be first applied for dimensionality reduction.

In face recognition, many techniques have been proposed (For a good review, see [2]). Among them, the most notable is a *two-stage* PCA+LDA approach [4,1]:

$$A = A_{\mathrm{LDA}} A_{\mathrm{PCA}}$$

Principal Component Analysis (PCA) is used to project images from the original *image space* into a *face-subspace*, where dimensionality is reduced and S_w is no longer degenerate, so that LDA can proceed without trouble. A potential problem is that the PCA criterion may not be compatible with the LDA criterion, thus the PCA step may discard dimensions that contain important discriminative information.

Chen et al. have recently proved that the null space of S_w contains the most discriminative information [2]. But, their approach fell short of making use of any information outside of that null space. In addition, heuristics are needed to extract a small number of features for image representation, so as to avoid computational problems associated with large scatter matrices.

In this paper, we present a direct, exact LDA algorithm for high dimensional data set. It accepts high dimensional data (such as raw images) as input, and optimizes Fisher's criterion directly, without any feature extraction or dimensionality reduction steps.

2. Direct LDA Solution

At the core of the direct LDA algorithm lies the idea of simultaneous diagonalization, the same as in the traditional LDA algorithm. As the name suggests, it tries to find a matrix that simultaneously diagonalizes both S_w and S_b:

$$AS_wA^T = I, \quad AS_bA^T = \Lambda$$

where Λ is a diagonal matrix with diagonal elements sorted in decreasing order. To reduce dimensionality to m, we simply pick the top m rows of A, which corresponds to the largest m diagonal elements in Λ. Details of the algorithm can be found in [3].

The key idea of our new algorithm is to discard the null space of S_b – which contains no useful information – rather than discarding the null space of

[1] Corresponding author. Email: hyu@cs.cmu.edu

Preprint submitted to Pattern Recognition Letters *29 September 2000*

S_w, which contains the most discriminative information. This can be achieved by diagonalizing S_b first and then diagonalizing S_w. The traditional procedure takes the reverse order. While both approaches produce the same result when S_w is not singular, the reversal in order makes a drastic difference for high dimensional data, where S_w is likely to be singular.

The new algorithm is outlined below. Figure 1 provides a conceptual overview of this algorithm. Computational issues will be discussed shortly after.

(1) Diagonalize S_b: find matrix V such that

$$V^T S_b V = \Lambda$$

where $V^T V = I$. Λ is a diagonal matrix sorted in decreasing order.

This can be done using the traditional eigen-analysis, i.e. each column of V is an eigenvector of S_b, and Λ contains all the eigenvalues. As S_b might be singular, some of the eigenvalues will be 0 (or close to 0). It is necessary to discard those eigenvalues and eigenvectors, as projection directions with a total scatter of 0 don't carry any discriminative power at all.

Let Y be the first m columns of V (an $n \times m$ matrix, n being the feature space dimensionality), now

$$Y^T S_b Y = D_b > 0$$

where D_b is the $m \times m$ principal sub-matrix of Λ.

(2) Let $Z = Y D_b^{-\frac{1}{2}}$,

$$(Y D_b^{-\frac{1}{2}})^T S_b (Y D_b^{-\frac{1}{2}}) = I \quad \Rightarrow \quad Z^T S_b Z = I$$

Thus, Z unitizes S_b, and reduces dimensionality from n to m.

Diagonalize $Z^T S_w Z$ by eigen-analysis:

$$U^T Z^T S_w Z U = D_w$$

where $U^T U = I$. D_w may contain 0s in its diagonal.

Since the objective is to maximize the ratio of total-scatter against within-class scatter, we can sort the diagonal elements of D_w and discard some eigenvalues in the high end, together with the corresponding eigenvectors. It is important to keep the dimensions with the smallest eigenvalues, especially 0s. This is exactly the reason why we started by diagonalizing S_b, rather than S_w. See Section 2.2 for more discussion.

(3) Let the LDA matrix

$$A = U^T Z^T$$

A diagonalizes both the numerator and the denominator in Fisher's criterion:

$$A S_w A^T = D_w, \quad A S_b A^T = I$$

(4) For classification purpose, notice that A already diagonalizes S_w, therefore the final transformation that spheres the data should be:

$$x^* \leftarrow D_w^{-\frac{1}{2}} A x$$

2.1. *Computational Considerations*

Although the scheme above gives an exact solution for Fisher's criterion, we haven't addressed the computational difficulty that both scatter matrices are too big to be held in memory, let alone their eigen-analysis.

Fortunately, the method presented by Turk and Pentland [5] for the eigenface problem is still applicable. The key observation is that scatter matrices can be represented in a way that both saves memory, and facilitates eigen-analysis. For example,

$$S_b = \sum_{i=1}^{J} n_i(\mu_i - \mu)(\mu_i - \mu)^T = \Phi_b \Phi_b^T \quad (n \times n)$$

where

$$\Phi_b = [\sqrt{n_1}(\mu_1 - \mu), \sqrt{n_2}(\mu_2 - \mu), \cdots] \quad (n \times J)$$

J is the number of classes, n_i is the number of training images for class i. Thus, instead of storing an $n \times n$ matrix, we need only to store Φ_b which is $n \times J$. The eigen-analysis is simplified by virtue of the following lemma:

Lemma 1 *For any $n \times m$ matrix L, mapping $x \rightarrow Lx$ is a one-to-one mapping that maps eigenvectors of $L^T L$ ($m \times m$) onto those of LL^T ($n \times n$).*

As $\Phi_b^T \Phi_b$ is an $J \times J$ matrix, eigen-analysis is affordable. In Step 2 of our algorithm, to compute eigenvalues for $Z^T S_w Z$, simply notice

$$S_w = \sum_i (x_i - \mu_{k_i})(x_i - \mu_{k_i})^T = \Phi_w \Phi_w^T$$

where

$$\Phi_w = [x_1 - \mu_{k_1}, x_2 - \mu_{k_2}, \cdots] \quad (n \times n_t)$$

n_t is the total number of images in the training set. Thus

$$Z^T S_w Z = Z^T \Phi_w \Phi_w^T Z = (\Phi_w^T Z)^T \Phi_w^T Z$$

We can again use the Lemma 1 to compute eigenvalues.

2.2. *Discussions*

Null space of S_w The traditional simultaneous diagonalization begins by diagonalizing S_w. If S_w is

$$\boxed{\frac{|S_b|}{|S_w|}} \xrightarrow[S_b]{\text{Diagonalize}} \boxed{\frac{|\Lambda|}{|S'_w|}} \xrightarrow[\text{\& Unitize } \Lambda]{\text{Discard 0s}} \boxed{\frac{|I|}{|S''_w|}} \xrightarrow[S''_w]{\text{Diagonalize}} \boxed{\frac{1}{|D_w|}} \xrightarrow[\text{(optional)}]{\text{Reduce dimensionality}}$$

Fig. 1. Thumbnail of the Direct LDA Algorithm

not degenerate, it gives the same result as our approach. If S_w is singular, however, the traditional approach runs into a dilemma: to proceed, it has to discard those eigenvalues equal to 0; but those discarded eigenvectors are the most important dimensions!

As Chen et al. pointed out [2], the null space of S_w [2] carries most of the discriminative information. More precisely, for a projection direction a, if $S_w a = 0$, and $S_b a \neq 0$, $\frac{a S_b a^T}{a S_w a^T}$ is maximized. The intuitive explanation is that, when projected onto direction a, within-class scatter is 0 but between-class scatter is not. Obviously perfect classification can be achieved in this direction.

Different from the algorithm proposed in [2], which operates solely in the null space, our algorithm can take advantage of all the information, both within and outside of S_w's null space. Our algorithm can still be used in cases where S_w is not singular, which is common in tasks like speech recognition.

Equivalence to PCA+LDA As Fukunaga pointed out [3], there are other variants of Fisher's criterion:

$$\arg\max_A \frac{|A^T S_t A|}{|A^T S_w A|} \quad or \quad \arg\max_A \frac{|A^T S_b A|}{|A^T S_t A|}$$

where $S_t = S_b + S_w$ is the *total scatter matrix.*

Interestingly, if we use the first variant (with S_t in the numerator), Step 1 of our algorithm becomes exactly PCA. Discarding S_t's eigenvectors with 0 eigenvalues reduces dimensionality, just as Belhumeur et al. proposed in their two-stage PCA+LDA method [1]. If their LDA step handled S_w's null space properly, the two approaches would give the same performance. In a sense our method can be called "unified PCA+LDA", since there is no separate PCA step. It not only leads to a clean presentation, but also results in an efficient implementation.

3. Face Recognition Experiments

We tested the direct LDA algorithm on face images from Olivetti-Oracle Research Lab (ORL, http://www.cam-orl.co.uk). The ORL dataset consists of 400 frontal faces: 10 tightly-cropped images of 40 individuals with variations in pose, illumination, facial expression (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). The size of each image is 92×112 pixels, with 256 grey levels per pixel.

Three sets of experiments are conducted. In all cases we randomly choose 5 images per person for training, the other 5 for testing. To reduce variation, each experiment is repeated at least 10 times.

Without dimensionality reduction in Step 2, average recognition accuracy is 90.8%. With dimensionality reduction, where everything outside of S_w's null space is discarded, average recognition accuracy becomes 86.6%. This verifies that while S_w's null space is important, discriminative information does exist outside of it.

4. Conclusions

In this paper, we proposed a direct LDA algorithm for high-dimensional data classification, with application to face recognition in particular. Since the number of samples is typically smaller than the dimensionality of the samples, both S_b and S_w are singular. By modifying the simultaneous diagonalization procedure, we are able to discard the null space of S_b – which carries no discriminative information – and to keep the null space of S_w, which is very important for classification. In addition, computational techniques are introduced to handle large scatter matrices efficiently. The result is a unified LDA algorithm that gives an exact solution to Fisher's criterion whether or not S_w is singular.

References

[1] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherface: Recognition using class specific linear projection. *PAMI*, 19(7):711–720, 1997.

[2] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu. A new lda-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33(10):1713–1726, Oct 2000.

[3] K. Fukunaga. *Introduction to Statistical Pattern Recognition (Second Edition).* New York: Academic Press, 1990.

[4] D. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *PAMI*, 18(8):831–836, August 1996.

[5] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):72–86, 1991.

[2] Null space of $S_w = \{x | S_w x = 0, x \in \mathrm{R}^n\}$.