Solving the Small Sample Size Problem of LDA

Rui Huang, Qingshan Liu, Hanqing Lu, Songde Ma
National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, P.O. Box 2728, Beijing, 100080, P.R. China
rhuang, qsliu, luhq, masd@nlpr.ia.ac.cn

Abstract

The small sample size problem is often encountered in pattern recognition. It results in the singularity of the within-class scatter matrix S_w in Linear Discriminant Analysis (LDA). Different methods have been proposed to solve this problem in face recognition literature. Some methods reduce the dimension of the original sample space and hence unavoidably remove the null space of S_w, which has been demonstrated to contain considerable discriminative information; whereas other methods suffer from the computational problem. In this paper, we propose a new method to make use of the null space of S_w effectively and solve the small sample size problem of LDA. We compare our method with several well-known methods, and demonstrate the efficiency of our method.

1. Introduction

Linear Discriminant Analysis (LDA) is used to seek a projection W, from the original sample space to a lower-dimensional space, which maximizes the between-class scatter while minimizing the within-class scatter. A typical way to achieve this is to maximize the ratio: $\frac{|W^T S_b W|}{|W^T S_w W|}$, where S_b is the between-class scatter matrix and S_w is the within-class scatter matrix. It has been proved that if S_w is a non-singular matrix then the ratio is maximized when the column vectors of W are the eigenvectors of $S_w^{-1} S_b$. Unfortunately, in many practical applications of pattern recognition, S_w is singular because the number of the samples is much smaller than the dimension of the sample space. This is called a small sample size problem [1]. Different methods have been proposed to solve this problem and applied to image retrieval, object and face recognition tasks. We compare these methods and give a new solution to this problem.

The rest of this paper is organized as follows: Section 2 reviews the related work on LDA-based methods for pattern recognition (mainly for face recognition); section 3 introduces one new method we proposed; the experiments are shown and discussed in section 4; and section 5 concludes this paper.

2. Related work

To simplify the discussion, some assumptions and definitions are given firstly.

Suppose the dimension of the original sample space is n, and a c-class problem is considered. The between-class scatter matrix S_b and the within-class scatter matrix S_w are defined as:

$$S_b = \frac{1}{N}\sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^T = \frac{1}{N}\Phi_b \Phi_b^T$$

$$S_w = \frac{1}{N}\sum_{i=1}^{c}\sum_{x \in C_i} (x - m_i)(x - m_i)^T = \frac{1}{N}\Phi_w \Phi_w^T$$

where N_i is the number of the samples in class C_i (i = 1, 2, …, c), $N = \sum_{i=1}^{c} N_i$ is the number of all the samples, $m_i = \frac{1}{N_i}\sum_{x \in C_i} x$ is the mean of the samples in class C_i, and $m = \frac{1}{N}\sum_{x} x$ is the mean of all the samples. Then the total scatter matrix or mixture scatter matrix S_t is defined by:

$$S_t = S_b + S_w = \frac{1}{N}\sum_{x} (x - m)(x - m)^T = \frac{1}{N}\Phi_t \Phi_t^T$$

which is also the covariance matrix of all the samples.

The goal of LDA is to find an optimal projection:

$$W = \arg\max_{W} \frac{|W^T S_b W|}{|W^T S_w W|}$$

It is easy to prove that the upper bounds of the rank of S_b, S_w and S_t are respectively $c-1$, $N-c$ and $N-1$, which are all much less than n in many practical problems, i.e. S_b, S_w and S_t are all usually singular in practice.

Null space (Kernel) of matrix A: $\{x | Ax = 0, x \in R^n\}$. Its dimension (**Nullity** of A) is: $n-$rank (A).

2.1. Regularization method

To deal with the singularity of S_w, a regularization method was mentioned in [2]. S_w can be slightly modified to S_w+KI, where K is a very small (relative to the eigenvalues of S_w) positive number such that S_w+KI is strictly positive definite. This is a pure LDA method without dimensionality reduction.

However, the computational complexity is very high to handle such a high-dimensional S_w.

2.2. Subspace method

Another kind of method to solve the small sample size problem is projecting the original samples to a lower-dimensional space to make the resulting within-class scatter matrix full-rank. Various subspaces have been used previously.

The most widely used subspace method [2, 3, 4] performs Principle Component Analysis (PCA) firstly to reduce the dimension of the samples from n to an intermediate dimension n_1, which must be not more than the rank of S_w (usually $N-c$) so as to make the resulting within-class scatter matrix full-rank. Then standard LDA is used to reduce the dimension further to n_2, which must be not more than $c-1$, for the rank of S_b is at most $c-1$.

Another novel method called Direct LDA [5] removes the null space of S_b firstly by doing eigen-analysis. Then a simultaneous diagonalization procedure is used to seek the optimal discriminant vectors in the subspace of S_b.

A more direct method is removing the null space of S_w firstly by doing eigen-analysis. Then standard LDA can be performed safely in the subspace of S_w.

Concerning the computational complexity, because calculating the eigenvalues and eigenvectors from an $n \times n$ matrix (e.g. S_t, S_b or S_w) is hard for typical sample sizes, and what we care about are only those eigenvectors corresponding to nonzero eigenvalues, a more efficient procedure [1] can work by firstly solving the eigenvalues and eigenvectors from a lower-dimensional matrix (e.g. $\frac{1}{N}\Phi_t^T\Phi_t$, $\frac{1}{N}\Phi_b^T\Phi_b$, or $\frac{1}{N}\Phi_w^T\Phi_w$).

These three methods do eigen-analysis on three different scatter matrices: S_t, S_b, or S_w respectively, but they are all based on dimensionality reduction and in fact remove the null space of S_w. It is notable that Direct LDA appears to avoid removing the null space of S_w, but cannot substantially avoid it. In fact, the rank of S_b is usually smaller than that of S_w, so the subspace that guarantees the full rank of S_b also guarantees the full rank of S_w. Therefore, removing the null space of S_b by dimensionality reduction would indirectly lead to the losing of the null space of S_w.

However, it was mentioned in [4] and investigated in detail in [6] that the optimal discriminant vectors of LDA could be derived from the null space (or kernel in [4]) of S_w. In fact, if a certain vector q belongs to the null space of S_w (i.e. $q^T S_w q = 0$), and also satisfies $q^T S_b q \neq 0$, then the ratio $\frac{|q^T S_b q|}{|q^T S_w q|}$ will definitely reach the maximum value. This means that the null space of S_w contains considerable discriminative information, whereas above subspace methods discard this information by removing the null space of S_w.

2.3. Null space method

In [6] an LDA-based method that makes use of the null space of S_w was proposed. All the samples are firstly projected onto the null space of S_w, where the within-class scatter is zero, and then the optimal discriminant vectors of LDA are those vectors that can maximize the between-class scatter. PCA is used to yield them.

Like the regularization method, the computational complexity of determining the null space of S_w is also very high because of the high dimension of S_w. So in [6] a pixel grouping method is used in advance to extract geometric features and to reduce the dimension of the samples, and then the null space LDA method is used in the feature space but not the original sample space.

3. Our work

We proposed a new method to solve the computational problem of the original null space LDA method. As mentioned above, if $q^T S_w q = 0$, and $q^T S_b q \neq 0$, then q is very useful for discrimination. But if $q^T S_w q = 0$, and $q^T S_b q = 0$ too, then q is not useful for discrimination. This means that not the whole null space of S_w is useful for discrimination. We can prove that the null space of S_t is the common null space of both S_b and S_w.

Proof:

It is known that: $S_t = S_b + S_w$.

Let Q be the null space of S_t, thus

$$Q^T S_t Q = 0$$
$$\Leftrightarrow Q^T (S_b + S_w) Q = 0$$
$$\Leftrightarrow Q^T S_b Q + Q^T S_w Q = 0$$
$$\Leftrightarrow Q^T S_b Q = 0 \quad \wedge \quad Q^T S_w Q = 0$$

(since S_b and S_w are positive semi-definite).

Q.E.D.

Therefore, the null space of S_t can be removed firstly by eigen-analysis without losing useful discriminative information. Only in the lower-dimensional projected space does the null space of the resulting within-class scatter matrix need to be determined. Through above procedure, a much smaller and equally useful subspace of the null space of S_w is found, which is then used to derive the optimal discriminant vectors of LDA. A similar concept was mentioned in [7], but they used an iteration algorithm, which also suffers from the computational problem.

We propose a new algorithm based on eigen-analysis and a procedure similar to simultaneous diagonalization. The whole algorithm and the computational considerations are described as follows:

1. Remove the null space of S_t.

This can be done by doing eigen-analysis on the $N \times N$ matrix $\frac{1}{N}\Phi_t^T \Phi_t$ instead of the $n \times n$ matrix S_t [1]. Let U be the matrix whose columns are all the eigenvectors of S_t corresponding to the nonzero eigenvalues, then we get:

$S'_w = U^T S_w U$ and $S'_b = U^T S_b U$.

2. Calculate the null space of S'_w.

After step 1, the dimension of S'_w is at most $N-1$, for the rank of S_t is at most $N-1$. It is now quite manageable to calculate the null space of S'_w by doing eigen-analysis again. The dimension of this null space (nullity of S'_w) is usually $c-1$, because the rank of S'_w is usually equal to that of S_w, which is usually $N-c$. Let Q be the null space of S'_w, then we get:

$S''_w = Q^T S'_w Q = (UQ)^T S_w (UQ) = 0$

and $S''_b = Q^T S'_b Q = (UQ)^T S_b (UQ)$.

UQ is a subspace of the whole null space of S_w, and is really useful for discrimination.

3. Remove the null space of S''_b if it exists, and reduce dimension further if necessary.

Do eigen-analysis on S''_b. Let V be the matrix whose columns are all the eigenvectors of S''_b corresponding to the nonzero eigenvalues or part of them associated with the largest eigenvalues (for further dimensionality reduction), then the final LDA projection is: $W = UQV$.

The last step is optional, because S''_b is usually full-rank. So the number of the optimal discriminant vectors was $c-1$, which coincide with the number of ideal features for classification [1].

4. Experiments

To demonstrate the efficiency of our method, extensive experiments are performed on different face data sets, and two of them are shown here. Besides our method, the method proposed in [2, 3, 4], called Fisherface method, and Direct LDA method proposed in [5] were tested. We do not test the regularization method because a detailed comparison between this method and Fisherface method has been demonstrated in [2]. We also do not list the test result of the original null space LDA method because it just has the same recognition accuracy as our method.

4.1. The ORL face database

There are 10 different images of 40 distinct subjects in the ORL face database. For some of the subjects, the images were taken at different times, varying lighting slightly, facial expressions (open/closed eyes, smiling/non-smiling) and facial details (glasses/no-glasses). All the subjects are in up-right, frontal position (with tolerance for some side movement). The size of each image is 92×112. Figure 1 shows 10 images of a subject.

Figure 1. Samples from the ORL database

We tested the recognition rates with different number of training samples. k (k = 2, 3, …, 9) images of each subject are randomly selected from the database for training and the remaining $10-k$ images of each subject for testing. For each value of k, at least 50 runs are performed with different random partition between training set and testing set, and table 1 shows the average recognition rates (%). No any pre-processing is done, and we choose 39 (i.e. $c-1$) as the final dimension. The detailed choice of intermediate dimension is described in references.

k	Fisherface	Direct LDA	Our Method
2	78.83	80.63	83.56
3	87.09	87.33	90.11
4	92.49	92.10	94.17
5	94.19	94.68	95.63
6	95.99	96.65	97.13
7	97.27	98.06	98.08
8	98.50	99.25	98.95
9	99.00	99.95	99.15

Table 1. Recognition rates on the ORL database

4.2. The FERET Database

To experiment on more challenging data, we have selected 70 subjects from the FERET database [8] with 6 up-right, frontal-view images of each subject. The number of subjects is more and the number of samples for each subject is less than the ORL database. The images were selected to bear with more differences in lighting, facial expressions and facial details. Figure 2 shows 2 subjects from the selected data set.

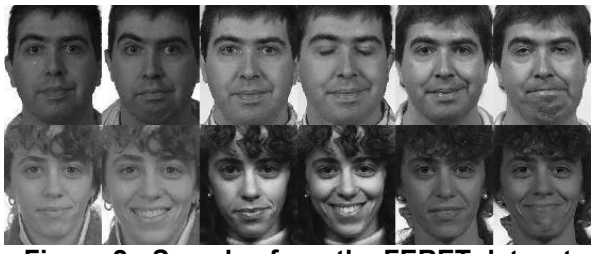

Figure 2. Samples from the FERET dataset

The eye locations are fixed by geometric normalization. The size of face images is normalized to 92×112 (to be consistent with the ORL database). No other preprocessing is done. Test process is the same as the ORL test. Here k is from 2 to 5, and final dimension is 69. The recognition rates (%) are shown in table 2.

k	Fisherface	Direct LDA	Our Method
2	56.04	63.25	75.60
3	76.95	76.71	86.47
4	87.23	88.30	93.07
5	94.80	94.71	97.64

Table 2. Recognition rates on the FERET dataset

4.3. Discussions

Some observations can be obtained from above experiments. Our algorithm outperforms Fisherface method and Direct LDA method on the whole, especially when the number of training samples is small, which is often the case in face recognition and other pattern recognition tasks. When the number of training samples is large, the recognition rates on the ORL database are very close. We conclude that is because there are too few testing samples.

As to the computational complexity, the most time-consuming procedure, eigen-analysis, is performed on two matrices ($c\times c$ and $c-1\times c-1$) in Direct LDA method, on two matrices ($N\times N$ and $N-1\times N-1$) in our method, and on three matrices (one of $N\times N$, and two of $N-c\times N-c$) in Fisherface method using simultaneous diagonalization [3]. Direct LDA method has the minimum computational complexity. The other two methods have similar complexity, which are also quite manageable.

The experimental results have shown that our method gets the best recognition rates with a relatively low computational complexity. We are currently performing various experiments on other areas.

5. Conclusions

In this paper, we propose a new method using a subspace of the null space of the within-class scatter matrix S_w to solve the small sample size problem of LDA. Knowing that the null space of S_w contains considerable discriminative information but it is difficult to determine and redundant to use, we firstly remove the null space of S_t, which has been proved to be the common null space of both S_b and S_w, and useless for discrimination. Then in the lower-dimensional projected space, the null space of the resulting within-class scatter matrix is calculated. This lower-dimensional null space, combined with the previous projection, represents a subspace of the whole null space of S_w, and is really useful for discrimination. The optimal discriminant vectors of LDA are derived from it. The efficiency of our method is verified by extensive experiments on face recognition.

Acknowledgements: Portions of the research in this paper use the *FERET* database of facial images collected under the *FERET* program, and the *ORL* face database from the Olivetti Research Laboratory in Cambridge, UK.

References

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.

[2] W. Zhao, R. Chellappa, and P.J. Phillips, "Subspace Linear Discriminant Analysis for Face Recognition", Technical Report CAR-TR-914, Center for Automation Research, University of Maryland, 1999.

[3] D.L. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 18, No. 8, 1996.

[4] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, 1997.

[5] J. Yang, Y. Yu, and W. Kunz, "An Efficient LDA Algorithm for Face Recognition", *Proceedings of the Sixth International Conference on Control, Automation, Robotics and Vision (ICARCV2000)*, 2000.

[6] L.F. Chen, H.Y.M. Liao, J.C. Lin, M.T. Ko, and G.J. Yu, "A New LDA-based Face Recognition System Which Can Solve the Small Sample Size Problem", *Pattern Recognition*, Vol. 33, No. 10, 2000.

[7] Y. Guo, X. Huang, and J. Yang, "A Novel Algorithm Solving Fisher Optimal Discriminant Vector and Facial Recognition"(In Chinese), *Journal of Image and Graphics*, Vol. 4, No. 2, 1999.

[8] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms", *Image and Vision Computing*, Vol. 16, No. 5, pp. 295-306, 1998.