EXTRACTING THE OPTIMAL DIMENSIONALITY FOR DISCRIMINANT ANALYSIS

Feiping Nie, Shiming Xiang, Yangqiu Song and Changshui Zhang

State Key Laboratory of Intelligent Technology and Systems
Department of Automation, Tsinghua University, Beijing 100080, China

ABSTRACT

For classification task, supervised dimensionality reduction is a very important method when facing with high-dimensional data. Linear Discriminant Analysis(LDA) is one of the most popular method for supervised dimensionality reduction. However, LDA suffers from the singularity problem, which makes it hard to work. Another problem is the determination of optimal dimensionality for discriminant analysis, which is an important issue but often been neglected previously. In this paper, we propose a new algorithm to address these two problems. Experiments show the effectiveness of our method and demonstrate much higher performance in comparison to LDA.

Index Terms— optimal dimensionality, supervised dimensionality reduction, linear discriminant analysis, singularity problem, image recognition

1. INTRODUCTION

Dimensionality reduction is an important method when facing with high-dimensional data, and many supervised dimensionality reduction algorithms have been proposed for the purpose of classification task. Among those supervised algorithms, Linear Discriminant Analysis(LDA) is one of the most popular one. It has been successfully applied in many classification task such as face recognition. However, there exist several drawbacks in LDA. Firstly, it suffers from the small sample size(SSS) problem when dealing with high dimensional data. In this case, the within-class scatter matrix S_w may become singular, which makes LDA difficult to work. Many approaches have been proposed to address this problem[1, 2, 3]. However, all these variants of LDA discard a subspace and some important discriminative information may be lost.

Another drawback of LDA is that the number of available projection directions in LDA is smaller than the class number [4] and it is insufficient for some complex problems.Moreover, based on the criterion of LDA, one can not determine the optimal dimensionality to be reduced since the optimal value of the criterion is monotonic with respect to projection dimensionality.

How to select a suitable dimensionality for discriminant analysis? This important issue was often neglected previously. In this paper, we tend to solve this problem. To this end, we propose a new criterion, under which the optimal value is not monotonic with respect to projection dimensionality. Then we can extract the optimal dimensionality for discriminant analysis based on this criterion. Simultaneously, the singularity problem in LDA does not occur naturally.

Using the kernel trick, our method can also be easily extended to the nonlinear case.

The rest of this paper is organized as follows: A brief view and analysis of LDA is present in Section 2. In Section 3, we propose a new algorithm to solve these problems. A nonlinear extension of our method is given using the kernel trick, which is described in Section 4. In Section 5, two experiments on image recognition are presented to demonstrate the effectiveness of our method. Finally, we give the conclusions in Section 6.

2. REVIEW OF LINEAR DISCRIMINANT ANALYSIS

Let $\boldsymbol{x}_i \in \mathbb{R}^d (i = 1, 2, ..., n)$ be d-dimensional data and $l_i \in \{1, 2, ..., c\}$ be associated class labels, where n is the number of data and c is the number of classes. Let n_i be the number of data in the class i.

LDA is to learn a linear transformation$\mathbf{W} : \mathbb{R}^d \to \mathbb{R}^m$, and $\mathbf{W} \in \mathbb{R}^{d\times m}$. Then the original high-dimensional data $\boldsymbol{x}$ is transformed into a low-dimensional vector:

$$\boldsymbol{y} = \mathbf{W}^T \boldsymbol{x} \tag{1}$$

With the projection matrix $\mathbf{W}$, LDA tries to maximize the between-class scatter, while minimizing the within-class scatter. The within-class scatter matrix S_w and the between-class scatter matrix S_b are defined as

$$\mathbf{S}_w = \sum_{i=1}^{c} \sum_{j:l_j=i} (\boldsymbol{x}_j - \boldsymbol{m}_i)(\boldsymbol{x}_j - \boldsymbol{m}_i)^T \tag{2}$$

$$\mathbf{S}_b = \sum_{i=1}^{c} n_i(\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^T \tag{3}$$

where $\boldsymbol{m}_i (i = 1, 2, ..., c)$ is the mean of the samples in class

This paper is Funded by Basic Research Foundation of Tsinghua National Laboratory for Information Science and Technology (TNList).

i and $\boldsymbol{m}$ is the mean of all the samples:

$$\boldsymbol{m}_i = \frac{1}{n_i} \sum_{j:l_j=i} \boldsymbol{x}_j \tag{4}$$

$$\boldsymbol{m} = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{x}_j \tag{5}$$

The projection matrix $\mathbf{W}^*$ in LDA is learned by solving the following optimization problem:

$$\mathbf{W}^* = \arg \max_{\mathbf{W}\in\mathbb{R}^{d\times m}} tr\left((\mathbf{W}^T\mathbf{S}_w\mathbf{W})^{-1}\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right) \tag{6}$$

where $tr(\cdot)$ denotes the trace operator. It has been known that the solution of this optimization problem is the m largest eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$, and the optimal value is $\sum_{i=1}^{m}\lambda_i$, where $\lambda_i(i = 1, 2, ..., m)$ are the first m largest eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b$ and m is the projection dimensionality[5].

From the solution we can directly see the two limitations of LDA. First, when $\mathbf{S}_w$ is singular, it cannot be solved numerically. Second, the optimal value monotonously increase when the projection dimensionality m increase. Therefore, one cannot determine the optimal dimensionality for discriminant analysis. In the next section, we propose a new algorithm to solve this two problems.

3. OPTIMAL DIMENSIONALITY DISCRIMINANT ANALYSIS

Similar to LDA, our goal is also to maximize the between-class scatter, while minimizing the within-class scatter. To avoid the singularity problem in LDA, we introduce another criterion here, i.e. we use the difference form other than the quotient form to formulate the criterion.

We further add a constraint $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ to avoid trivial solution, where $\mathbf{I}$ is $m \times m$ identity matrix. Then the new criterion can be written as

$$\mathbf{W}^* = \arg \max_{\substack{\mathbf{W}\in\mathbb{R}^{d\times m}\\ \mathbf{W}^T\mathbf{W}=\mathbf{I}}} tr\left(\mathbf{W}^T(\mathbf{S}_b - \gamma\mathbf{S}_w)\mathbf{W}\right) \tag{7}$$

When the projection dimensionality is d, namely, the data is in the original space without dimensionality reduction, we let the value of this criterion be equal to zero so that the optimal value could reach the highest point in the reduced subspace. Thus, we have

$$\begin{aligned}
&\max_{\substack{\mathbf{W}\in\mathbb{R}^{d\times d}\\ \mathbf{W}^T\mathbf{W}=\mathbf{I}}} tr\left(\mathbf{W}^T(\mathbf{S}_b - \gamma\mathbf{S}_w)\mathbf{W}\right) = 0\\
&\Rightarrow tr\left(\mathbf{S}_b - \gamma\mathbf{S}_w\right) = 0\\
&\Rightarrow \gamma = \frac{tr\mathbf{S}_b}{tr\mathbf{S}_w}
\end{aligned}$$

Our goal is to extract the optimal dimensionality so that the optimal value reaches the maximum. Then the optimization problem can be formulated as follows:

$$\mathbf{W}^* = \arg \max_{m\in[1,...,d]} \max_{\substack{\mathbf{W}\in\mathbb{R}^{d\times m}\\ \mathbf{W}^T\mathbf{W}=\mathbf{I}}} tr\left(\mathbf{W}^T(\mathbf{S}_b - \frac{tr\mathbf{S}_b}{tr\mathbf{S}_w}\mathbf{S}_w)\mathbf{W}\right) \tag{8}$$

Denote $\mathbf{W} \in \mathbb{R}^{d\times m}$ by $\mathbf{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_m]$, where $\boldsymbol{w}_i(i = 1, 2, ...m)$ are d-dimensional column vectors. Suppose the value of m is given, according to Ky Fan's Theorem [6], when $\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_m$ are the first m largest eigenvectors of $\mathbf{S}_b - \frac{tr\mathbf{S}_b}{tr\mathbf{S}_w}\mathbf{S}_w$, the optimal value of the above optimization problem is $\sum_{i=1}^{m}\lambda_i$, where $\lambda_i(i = 1, 2, ..., m)$ are the first m largest eigenvalues of $\mathbf{S}_b - \frac{tr\mathbf{S}_b}{tr\mathbf{S}_w}\mathbf{S}_w$. Thus, when m is equal to the number of positive eigenvalues of $\mathbf{S}_b - \frac{tr\mathbf{S}_b}{tr\mathbf{S}_w}\mathbf{S}_w$, the optimal value reaches the maximum. Therefore, the optimal solution of the optimization problem in Eq. (8) can be explicitly calculated by eigenvalue decomposition.

The algorithm is described in Table 1. From the algorithm we can see, the singularity problem in LDA does not exist in it naturally.

```
0. Preprocessing: eliminate the null space of the covariance matrix of data, and obtain new data $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$, where $rank(\mathbf{X}) = d$
1. Input:
    $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$
2. calculate $\mathbf{S}_w$ and $\mathbf{S}_b$ according to Eq. (2) and Eq. (3)
3. calculate the eigenvalues and the corresponding eigenvectors of $\mathbf{S}_b - \frac{tr\mathbf{S}_b}{tr\mathbf{S}_w}\mathbf{S}_w$
4. Select the $m$ largest eigenvectors to form $\mathbf{W}$, where $m$ is equal to the number of positive eigenvalues of $\mathbf{W}$.
5. Output:
    $\boldsymbol{y} = \mathbf{W}^T\boldsymbol{x}$, where $\mathbf{W} \in \mathbb{R}^{d\times m}$ and $\mathbf{W}^T\mathbf{W} = \mathbf{I}$.
```

Table 1. *Algorithm of extracting the optimal dimensionality for discriminant analysis*

4. KERNELIZATION FOR NONLINEAR EXTENSION

In this section we show that, using the kernel trick, it can be easy to extend the linear projections of our algorithm to the nonlinear case.

Suppose the data is mapped from the original input space to a higher dimensional Hilbert space with a nonlinear mapping $\phi : x \rightarrow \mathcal{F}$. If the algorithm only needs to calculate the inner product of data pairs, using a kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i) \cdot \phi(\boldsymbol{x}_j)$, it can be performed in the new feature space without the explicit function of mapping. An popular choice of the kernel function is the Gaussian kernel

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right) \tag{9}$$

For convenience, we denote the data matrix in the feature space by $\mathbf{X} = [\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2), ..., \phi(\boldsymbol{x}_n)]$. From the view of graph [7], $\mathbf{S}_w$ and $\mathbf{S}_b$ can be reformulated as

$$\mathbf{S}_w = \mathbf{X}\mathbf{L}_w\mathbf{X}^T \tag{10}$$

$$\mathbf{S}_b = \mathbf{X}\mathbf{L}_b\mathbf{X}^T \tag{11}$$

where $\mathbf{L}_w$ and $\mathbf{L}_b$ are the Laplacian matrix on graph.

A Laplacian matrix is defined as $\mathbf{D} - \mathbf{S}$, where $\mathbf{S}$ is the similarity matrix of graph, and $\mathbf{D}$ is a diagonal matrix, whose entries are column(or row since $\mathbf{S}$ is symmetric) sums of $\mathbf{S}$, $\mathbf{D}_{ii} = \sum_j \mathbf{S}_{ji}$. For $\mathbf{L}_w$, the entities of the similarity matrix are defined as follows

$$\mathbf{S}_{ij} = \begin{cases} \frac{1}{n_k} & \boldsymbol{x}_i \text{ } and \text{ } \boldsymbol{x}_j \text{ } belong \text{ } to \text{ } class \text{ } k \\ 0 & otherwise \end{cases} \tag{12}$$

For $\mathbf{L}_b$, the entities of the similarity matrix are defined as follows

$$\mathbf{S}_{ij} = \begin{cases} \frac{1}{n} - \frac{1}{n_k} & \boldsymbol{x}_i \text{ } and \text{ } \boldsymbol{x}_j \text{ } belong \text{ } to \text{ } class \text{ } k \\ \frac{1}{n} & otherwise \end{cases} \tag{13}$$

Using the kernel function, the kernel matrix $\mathbf{K} = \mathbf{X}^T\mathbf{X}$ can be calculated. According to Eq. (10) and Eq. (11), we have

$$tr(\mathbf{S}_w) = tr(\mathbf{X}\mathbf{L}_w\mathbf{X}^T) = tr(\mathbf{L}_w\mathbf{X}^T\mathbf{X}) = tr(\mathbf{L}_w\mathbf{K}) \tag{14}$$

$$tr(\mathbf{S}_b) = tr(\mathbf{X}\mathbf{L}_b\mathbf{X}^T) = tr(\mathbf{L}_b\mathbf{X}^T\mathbf{X}) = tr(\mathbf{L}_b\mathbf{K}) \tag{15}$$

In the feature space, $\mathbf{W}$ can be expressed as $\mathbf{W} = \mathbf{X}\boldsymbol{\alpha}$, then

$$\mathbf{W}^T(\mathbf{S}_b - \frac{tr\mathbf{S}_b}{tr\mathbf{S}_w}\mathbf{S}_w)\mathbf{W} = \boldsymbol{\alpha}^T\mathbf{K}\mathbf{L}\mathbf{K}\boldsymbol{\alpha} \tag{16}$$

$$\mathbf{W}^T\mathbf{W} = \boldsymbol{\alpha}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha} = \boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} \tag{17}$$

where $\mathbf{L} = \mathbf{L}_b - \frac{tr\mathbf{S}_b}{tr\mathbf{S}_w}\mathbf{L}_w$. Then the optimization problem in Eq. (8) in the feature space can be rewritten as

$$\boldsymbol{\alpha}^* = \arg \max_{m \in [1,...,n]} \max_{\substack{\boldsymbol{\alpha} \in \mathbb{R}^{n \times m} \\ \boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} = \mathbf{I}}} tr\left(\boldsymbol{\alpha}^T\mathbf{K}\mathbf{L}\mathbf{K}\boldsymbol{\alpha}\right) \tag{18}$$

The solution of Eq. (18) can be obtained by solving the generalized eigenvalue decomposition problem as

$$\mathbf{K}\mathbf{L}\mathbf{K}\boldsymbol{\alpha}_i = \mu\mathbf{K}\boldsymbol{\alpha}_i \tag{19}$$

$\boldsymbol{\alpha}_i$ should be resized as $\frac{1}{\sqrt{\boldsymbol{\alpha}_i^T\mathbf{K}\boldsymbol{\alpha}_i}}\boldsymbol{\alpha}_i$ to satisfy the constraint of $\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} = \mathbf{I}$.

Finally we get the projection matrix $\mathbf{W}^* = \mathbf{X}\boldsymbol{\alpha}^*$, where $\boldsymbol{\alpha}^* \in \mathbb{R}^{n \times m}$ and m is the number of positive eigenvalues of $\mathbf{KLK}$.

With the projection matrix $\mathbf{W}^*$, every data point $\boldsymbol{x}$ can be projected as $\boldsymbol{y} = \mathbf{W}^{*T}\boldsymbol{x} = \boldsymbol{\alpha}^{*T}\mathbf{X}^T\boldsymbol{x}$, where $\mathbf{X}^T\boldsymbol{x}$ can be calculated by the kernel function.

number	method	Accuracy(%)	Std. Dev.(%)
3	LDA	86.2	2.6
	NLDA	90.1	2.1
	DLDA	86.1	2.3
	ODLDA	**91.0**	2.2
4	LDA	89.4	2.2
	NLDA	92.8	1.6
	DLDA	91.2	1.8
	ODLDA	**94.2**	1.6
5	LDA	90.6	2.3
	NLDA	94.3	1.8
	DLDA	93.7	1.8
	ODLDA	**96.0**	1.5
6	LDA	91.5	2.0
	NLDA	94.7	1.6
	DLDA	95.8	1.4
	ODLDA	**97.0**	1.3

Table 2. *Experimental results on the AT&T face database.*

5. EXPERIMENTS

We evaluated our algorithm (denoted as ODLDA here) on two popular databases, and compared it with LDA and its two variants, null space LDA(denoted as NLDA)[2] and direct LDA(denoted as DLDA)[3]. As it is very hard to determine a suitable parameter in Eq. (9) to achieve a good performance, we do not evaluate the kernel version of our algorithm.

PCA is used as a preprocessing step to eliminate the null space of data covariance matrix $\mathbf{S}_t$. For LDA, due to the singularity problem in it, we further reduce the dimension of data such that the within-class scatter matrix $\mathbf{S}_w$ is nonsingular.

In each experiment, we randomly select several samples per class for training and the remaining samples are used for testing. the average results and standard deviations are reported over 50 random splits. The classification is based on k-nearest neighbor classifier($k = 1$ in these experiments).

The experimental results are reported in Table 2 and Table 3. In the following we describe the details of each experiment.

5.1. Face recognition

The AT&T face database (formerly the ORL database) includes 40 distinct individuals and each individual has 10 different images. Some images were taken at different times, and have variations [8] including expression and facial details. Each image in the database is of size 112×92 and with 256 gray-levels.

In this experiment, each image is down-sampled to the size of 28×23 to save the computation time. we randomly select 3,4,5 or 6 samples per class for training and the remaining samples for testing. As can been seen in Table 2, the results of our method is much better than those of LDA

number	method	Accuracy(%)	Std. Dev.(%)
4	LDA	77.3	2.6
	NLDA	81.8	3.0
	DLDA	80.2	2.3
	ODLDA	**83.6**	2.7
6	LDA	82.1	1.9
	NLDA	86.3	1.8
	DLDA	85.7	1.8
	ODLDA	**88.7**	1.7
8	LDA	84.8	1.3
	NLDA	89.0	1.4
	DLDA	89.7	1.6
	ODLDA	**91.7**	1.5
12	LDA	88.0	1.4
	NLDA	91.7	1.3
	DLDA	94.0	1.2
	ODLDA	**95.3**	1.2

Table 3. *Experimental results on the COIL-20 object database*.

whether in terms of accuracy or stability, and also outperform those of the two popular variants of LDA. It is interesting to note that the optimal dimensionality found by our method is just $c-1$, where c is the class number. That is to say, the number of positive eigenvalues of $\mathbf{S}_b - \frac{tr\mathbf{S}_b}{tr\mathbf{S}_w}\mathbf{S}_w$ is just equal to the maximum number that LDA can obtain. It illustrates the facts that the projection dimensionality in LDA should not less than $c-1$, and that when the projection dimensionality is $c-1$, LDA would reach its maximum performance.

5.2. Object recognition

The COIL-20 database [9] consists of images of 20 objects viewed from varying angles at the interval of five degrees, resulting in 72 images per object.

In this experiment, each image is down-sampled to the size of 32 × 32 to save the computation time. we randomly select 4,6,8 or 12 samples per class for training and the remaining samples for testing.

Similar to the face recognition experiment, the results of our method are much better than those of LDA, and the optimal dimensionality found by our method is also just equal to $c-1$. One possible reason of our method outperforming LDA and its variants may be that our method does not exist the singularity problem. Thus some important discriminant information would not lose during the process of dimensionality reduction.

6. CONCLUSIONS

The singularity problem is one of the most serious drawbacks in LDA which makes it hard to work. Although many methods have been proposed to solve this problem, the intrinsic problem in theory still exists. Another problem is the determination of the optimal dimensionality for discriminant analysis. Traditional LDA cannot determine the optimal dimensionality since the optimal value of the criterion in LDA is monotonic with respect to projection dimensionality, and this important issue has often been neglected previously. In this paper, we propose a new algorithm to address the two problems. The optimal dimensionality can be determined based on a new criterion, and the singularity problem does not occur intrinsically in the new algorithm. Two image databases have been used to validate our method. The experimental results show that our method is effective and the performances are much higher in comparison to LDA.

7. REFERENCES

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, July 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*, vol. 33, no. 10, pp. 1713–1726, October 2000.

[3] H. Yu and J. Yang, “A direct lda algorithm for high-dimensional data - with application to face recognition,” *Pattern Recognition*, vol. 34, pp. 2067–2070, 2001.

[4] Richard O. Duda., Peter E. Hart., and David G. Stork, *Pattern Classification*, Wiley-Interscience, Hoboken, NJ, 2000.

[5] Keinosuke Fukunaga, *Introduction to Statistical Pattern Recognition,Second Edition.*, Academic Press, Boston, MA, 1990.

[6] Gene H. Golub and Charles F. van Loan, *Matrix Computations, 3rd Edition.*, The Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[7] X. F. He, S. C. Yan, Y. X. Hu, P. Niyogi, and H. J. Zhang, “Face recognition using laplacianfaces,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, March 2005.

[8] F. S. Samaria and A. C. Harter, “Parameterisation of a stochastic model for human face identification,” in *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision*, 1994, pp. 138–142.

[9] S. A. Nene, S. K. Nayar, and H. Murase, *Columbia object image library (COIL-20), Technical Report CUCS-005-96*, Columbia University, 1996.