A Novel Method to Solve the Separation Problem of LDA

Meng Zhang, Wei Li, Bo Zhang

A novel method to solve the separation problem of LDA

Meng Zhang[1], Wei Li[1], and Bo Zhang[1,2]

[1] School of Mechatronic Engineering, China University of Mining and Technology, Xuzhou 221116, People's Republic of China
[2] School of Computer Science And Technology, China University of Mining and Technology
Xuzhou 221116, People's Republic of China
{zhangmeng,liwei_cmee,zbcumt}@163.com

Abstract. Linear discriminant analysis (LDA) is one of the most classical linear projection techniques for feature extraction, widely used in kinds of fields. Classical LDA is contributed to finding an optimal projection subspace that can maximize the between-class scatter and minimize the average within-class scatter of each class. However, the class separation problem always exists and classical LDA can not guarantee that the within-class scatter of each class get its minimum. In this paper, we proposed the *k-classifiers* method, which can reduce every within-class scatter of classes respectively and alleviate the class separation problem. This method will be applied in LDA and Norm LDA and achieve significant improvement. Extensive experiments performed on MNIST data sets demonstrate the effectiveness of *k-classifiers*.

Keywords: Linear Discriminant Analysis · Class Separation Problem · Within-class Scatter.

1 Introduction

Linear discriminant analysis (LDA) is one of the most popular linear dimension reduction (LDR) and feature extraction methods, which has been widely used in kinds of fields, such as face recognition, cancer classification and text document classification [9]. It is a supervised learning technique that finds an orientation W to project the feature vectors from the original sample space to a lower space in such a way that maximizes the between-class scatter and minimizes the within-class scatter simultaneously.

LDA was first proposed by Fisher (FLDA) [2] to solve the binary classification problem and then was generalized by Rao [6] to multiple classes. Although LDA is a classical LDR method, it still suffers from the class separation problem, close class pairs tend to mix up in the subspace. For a K-class problem, the Fisher criterion is actually decomposed into pairwise Fisher criteria under certain assumptions [5]. Conventional LDA seeks to maximize the average pairwise distance between class means and minimize the average within-class pairwise

distance over all classes. It is, in fact, desirable for every pairwise distance between two class means to be as large as possible and every within-class pairwise distance to be as small as possible. There are three main kinds of methods used to relieve the class separation problem. First, it is the optimal choice to design a Bayes optimal criterion for general multi-class discriminative dimension reduction[7, 3, 12]. However, this kind of method is quite difficult, as the Bayes error cannot be expressed analytically [3]. Second, many weighting methods have been proposed. In these methods, approximate weights are introduced into FLDA [5, 10, 1]. Whereas, the above weights methods cannot address the problem thoroughly. Lastly, the max-min methods have also been applied to solve the Class separation problem. These methods focus on maximizing the minimum pairwise between-class distance of all classes in the projected subspace [1, 13, 8].

The methods above almost all focus on how to make the pairwise distance larger and larger, and they all ignore reducing each within-class pairwise distance except WLDA [13]. However, for WLDA, only the maximum within-class scatter is minimized. We propose a novel method called *k-classifiers* to reduce every within-class scatter respectively and the method is applied in LDA and Norm LDA to improve their corresponding classification performance in this paper. For LDA, *k-classifiers* method can make every within-class of each class smaller as far as possible. For Norm LDA, *k-classifiers* method can make the scatter of orientation in maximum scatter smaller as far as possible. The difference of *k-classifiers* method in LDA and Norm LDA depends on its own geometric meaning. Norm LDA is an LDA method which based on different criterion refer to section 2.2.

The rest of the paper is organized as follows: Section 2 introduces briefly FLDA and Norm LDA. The idea of *k-classifiers* and the corresponding classification strategy are presented in section 3. The experiments are presented in Section 4. Finally, we give conclusions and future work in Section 5.

2 Related work

In this section, we briefly review two supervised dimensionality reduction methods, i.e., LDA and Norm LDA, which is the basis of the proposed method.

2.1 Classical LDA

Given a set of data containing C classes $\{Z_i\}_{i=1}^{C}$, with each class consisting of a number of samples:$Z_i = \{z_{ij}\}_{j=1}^{C_i}$, a total of $N = \sum_{i=1}^{C} N_i$ samples are available in the set. Each sample is represented as a column vector of length n, i.e. $z_{ij} \in R^n$, where R^n denotes the n-dimensional real space. We can define the within-class scatter S_w and between-class scatter S_b as follow:

$$S_w = \frac{1}{N} \sum_{i=1}^{C} \sum_{x \in Z_i} (x - m_i)(x - m_i)^T, \tag{1}$$

$$S_b = \frac{1}{N}\sum_{i=1}^{C}(m_i - m)(m_i - m)^T, \tag{2}$$

Where $m_i = \frac{1}{N_i}\sum_{x\in Z_i} x$ is the mean of the samples in class Z_i, and $m = \frac{1}{N}\sum_{x\in Z} x$ is the mean of all the samples.

For multi-class problem, based on maximizing the between-class scatter and minimizing the within-class scatter simultaneously, the trace ratio criterion is proposed naturally:

$$W = \arg\max_W \frac{trace(W^T S_b W)}{trace(W^T S_w W)}. \tag{3}$$

In fact, there does not exist a closed-form solution for the trace ratio criterion [11]. For easy to solve, a suboptimal substitute of the trace ratio criterion has been proposed, which called determinant ratio (ratio trace):

$$W = \arg\max_W \frac{|W^T S_b W|}{|W^T S_w W|}. \tag{4}$$

Solving the above criterion with the Lagrange equation, we can find that the basis vectors W correspond to the first M eigenvectors with the largest eigenvalues of $(S_w^{-1}S_b)$, when S_w is non-singular.

2.2 Norm LDA

As we all know, the conventional criterion or the criterion having the similar geometric meaning with conventional criterion can't always get the optimal performance in all database. Since S_w and S_b are positive semi-definite, we can always find Q and R such that $S_w = QQ^T$ and $S_b = RR^T$. A series of objective functions can be represented as:

$$J(W) = \arg\max_W \frac{\parallel (W^T R)^T \parallel}{\parallel (W^T Q)^T \parallel}, \tag{5}$$

Where $\parallel \cdot \parallel$ is a sub-multiplicative and unitary invariant matrix norm, i.e. $\parallel AB \parallel \leq \parallel A \parallel \parallel B \parallel$ with A and B being any compatible matrix, and $\parallel AB \parallel = \parallel B \parallel$ with B being any unitary matrix.

By using the F-norm, the objective (5) is equivalent to:

$$\frac{trace(W^T S_b W)}{trace(W^T S_w W)} = \frac{\parallel (W^T R)^T \parallel}{\parallel (W^T Q)^T \parallel}, \tag{6}$$

Which is the trace ratio of WS_bW^T and WS_wW^T.

By using the 2-norm, the objective (5) becomes:

$$J(W) = \arg\max_W \frac{\parallel (W^T R)^T \parallel_2}{\parallel (W^T Q)^T \parallel_2}, \tag{7}$$

Which is the ratio between the largest eigenvalue of WS_bW^T and WS_wW^T.

We can also define the objective function by using mixed norms, i.e.

$$J(W) = \arg\max_{W} \frac{\| (W^T R)^T \|_F}{\| (W^T Q)^T \|_2}. \tag{8}$$

Q can be decomposed by singular value decomposition (SVD) as:

$$Q = U\Sigma V^T, \tag{9}$$

where $UU^T = I_n \times n$, $VV^T = I_n \times n$.

Without loss of generality, we set $W = U\Sigma^- \tilde{W}$. Then we can get the unified analytical solution to the objective function (7) and (8), $W = U\Sigma^-$. However, it is not the solution to the objective function (6). We should emphasize that this unified analytical solution is only a projection and will not reduce the dimension of the feature space, and thus, we should conduct PCA to reduce dimensionality before using this method when the SSS problem occurs. According to the geometric meanings of SVD and eigenvalue, the distribution of discriminant information in each feature generated by Norm LDA is more uniform than in LDA. i.e. $W \in R^n \times n$, and any $W \neq U\Sigma^-$ results in

$$J(W) \leq J(U\Sigma^-). \tag{10}$$

As we know many LDA methods reduce the dimension of the feature space through the linear projection. Nevertheless, it is apparent that any W with $m < n$ cannot deliver a better result than $W = U\Sigma^-$ in sense of Eqs. (7) and (8).

3 Method

As formulated in Eq.(3), LDA simultaneously seeks to maximize the average of between-class scatter of each two classes and minimize the average within-class scatter of each all classes. However, the Fisher criterion cannot guarantee class separation since within-class scatter matrixes of each class are different. In this section, *k-classifiers* method is proposed to ensure the every within-class scatter as smaller as possible by designing k criterion according to k classes.

3.1 k-classifiers method

Methods about LDA proposed in the above literatures all contain just one projection orientation. They cannot diminish every within-class scatter matrix as far as possible. When the *k-classifiers* method is applied into LDA, the k criterions can be got corresponding to k within-class scatter matrixes as follow:

$$F(W_i) = \arg\max_{W_i} \frac{|W_i^T S_b W_i|}{|W_i^T S_{wi} W_i|}, \tag{11}$$

where $i = 1, 2, ..., k$. Inspired by the solution in section 2.1, i-th ($1 \leq i \leq k$) criterion can be solved and we can find that when $S_w i$ is non-singular, the basis

vectors W_i correspond to the first M eigenvectors with the largest eigenvalues of $(S_w^{-1}S_b)$. According to the geometric meaning of the above criterion, we can find that for LDA, *k-classifiers* method can make the every within-class of each class smaller as far as possible.

When the *k-classifiers* is applied into Norm LDA, we can obtain k criterions as follow:

$$J(W_i) = \arg\max_{W_i} \frac{\parallel (W_i^T R)^T \parallel}{\parallel (W_i^T Q)^T \parallel}, \tag{12}$$

where $i = 1, 2, ..., k$. By the method proposed in section 2.2, we can solve the above k criterions. The i-th optimal projection matrixes can be calculated and $W_i = \Sigma_i^- U_i$

By the prove in the section 2.2, we can work out that any $W_i \neq \Sigma_i^- U_i$ results in $J(W_i) \leq J(U_i \Sigma_1^-)$. We can define k classifiers by the k optimal projection matrixes according to the geometric meaning of 2-Norm and the above criterion, it's obvious that for Norm LDA, *k-classifiers* method can make the scatter of orientation in maximum scatter smaller as far as possible.

Similarly, this method also can be applied to many other methods proposed in the literature.

3.2 Classification strategy

Assuming normal distribution for each class with the common covariance matrix, classification based on maximum likelihood estimation results in a nearest class centroid rule. Assuming equal prior for all classes for simplicity, a test point y is classified as class j if

$$\parallel W_j^T (y - c^j) \parallel_F^2 \tag{13}$$

is minimized over j=1,...,k. It can be shown as:

$$\arg\max_j \{\parallel W_j^T (y - c^j) \parallel_F^2\}, \tag{14}$$

where W_j is the optimal projection matrix corresponding to LDA and Norm LDA and c^j is described as follows, $i = 1, 2, ..., k$:

$$c^j = \frac{1}{N_i} \sum_{x \in z_i} W_i^T x. \tag{15}$$

Classifier can be built based on Eq.(14). The applications of *k-classifiers* method in LDA and Norm LDA are described as follows:

1. Calculate the k optimal projection places of LDA and Norm LDA from train data by the $(S_{wi}^- S_b)$ and $\Sigma_i^- U_i$ respectively.
2. Project i-th class into i-th place and the center of i-th class in i-th place is achieved, $i = 1, 2, ..., k$.
3. For a test point y, the j-th distance between y and the each j-th class in the corresponding place can be calculated by Eq. (13).
4. Classifiers can be built by the Eq. (14). If the j-th distance is the minimum of k distances, then the point y belongs to j-th class.

4 Experiment

In order to demonstrate the effectiveness of the proposed fault diagnosis method, the MNIST database is used.

4.1 Experiments on the MNIST database

The MNIST database [4] of handwritten digits is a widely known benchmark that consists of a training set of 60,000 examples, and a test set of 10,000 examples.

In this experiment, we conduct PCA firstly to reduce the dimension. Then we select the training set to calculate the optimal projection matrixes corresponding to four methods. The dimensionality of the subspace generated by LDA is at most $k-1$, which depends on the rank of the between-class scatter matrix. When the dimension reduced by PCA is less than nine-dimensional, LDA will not be used as a dimension reduction method, and it just as a projection method. Otherwise, the subspace generated by LDA will be a nine-dimensional space. Lastly, classifiers corresponding to four methods will be designed based on the methods proposed in section 2. We repeat the experiment 20 times, and the average classification accuracy rates have been shown in Fig.1.

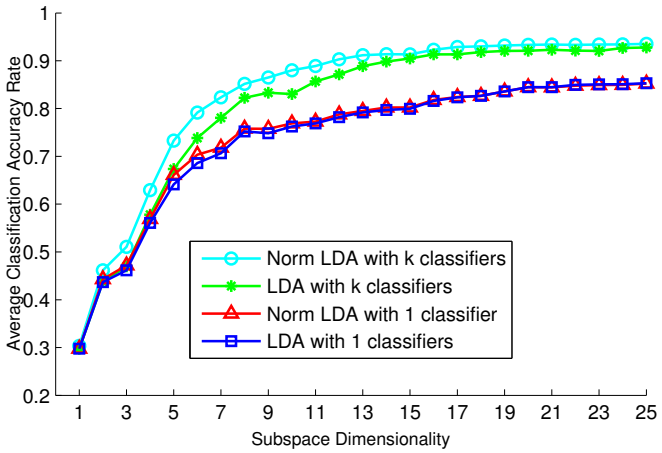

Fig. 1. Average classification accuracy rates on the MNIST database

Fig.1 shows that on the whole, the methods with *k-classifiers* obtain better classification results than the methods with 1 classifier. So, the application of *k-classifiers* is very successful and *k-classifiers* can actually alleviate the Class separation problem. We can conclude that decreasing the within-class scatter of each class by the *k-classifiers* method can actually improve the classification performance. It is also obvious that Norm LDA is a little better than LDA in both low and high dimension, under this kind of data distribution.

4.2 Experimental result Analysis

We summarize the observations from the above experiments and then make the analysis as follows:

(1) On the whole, the method of Norm LDA with *k-classifiers* outperforms the other three methods in the high dimension, and sometimes in the low dimension. LDA emphasizes the increase of determinant ratio between-class scatter and average within-class scatter of each class. Norm LDA devotes to increasing the ratio between the largest eigenvalue of between-class scatter and average within-class scatter of each class. For some datasets, Norm LDA is more suitable than LDA.

(2) The application of *k-classifiers* in Norm LDA is very successful. When the method of *k-classifiers* is applied in LDA, the method of LDA with *k-classifiers* is better than LDA with 1 classifier in both low and high dimensions. Norm LDA and LDA emphasize the equal decrease of within-class scatter of each class. The combination of *k-classifiers* and LDA, Norm LDA is devoting to the reduction of the within-class scatter of each class. As a result, on the whole, methods with *k-classifiers* can have a better performance.

5 Conclusions and future work

The class separation problem occurring in LDA has been continuously studied in recent years. Many ideas have been applied in LDA to improve its performance, including weights schemes, max-min, and Bayes optimal criterion. The application of weights schemes and max-min methods all ignored the importance of decreasing every within-class scatter. In this paper, we present a *k-classifiers* method to reduce every within-class pairwise. We also apply *k-classifiers* method into LDA and Norm LDA. Based on the MNIST handwriting database, we have demonstrated that the applications of *k-classifiers* method in LDA and Norm LDA are very successful.

There is still room to improve the classification performance. We can put the max-min ideas and weights schemes into our method to make every pairwise distance between two classes as larger as possible, and make within-class scatter of every class as smaller as possible.

6 Acknowledgments

The work was supported by China Postdoctoral Science Foundation(No. 2017M621862) and Jiangsu Planned Projects for Postdoctoral Research Funds(No.1701193B).

References

1. Bian, W., Tao, D.: Max-min distance analysis by using sequential sdp relaxation for dimension reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(5), 1037–1050 (2010)

2. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of eugenics **7**(2), 179–188 (1936)
3. Hamsici, O.C., Martinez, A.M.: Bayes optimality in linear discriminant analysis. IEEE transactions on pattern analysis and machine intelligence **30**(4), 647–657 (2008)
4. LeCun, Y., Cortes, C., Burges, C.: Mnist handwritten digit database (2010)
5. Loog, M., Duin, R.P.W., Haeb-Umbach, R.: Multiclass linear dimension reduction by weighted pairwise fisher criteria. IEEE transactions on pattern analysis and machine intelligence **23**(7), 762–766 (2001)
6. Rao, C.R.: The utilization of multiple measurements in problems of biological classification. Journal of the Royal Statistical Society. Series B (Methodological) **10**(2), 159–203 (1948)
7. Schervish, M.J.: Linear discrimination for three known normal populations. Journal of statistical planning and inference **10**(2), 167–175 (1984)
8. Shao, G., Sang, N.: Max–min distance analysis by making a uniform distribution of class centers for dimensionality reduction. Neurocomputing **143**, 208–221 (2014)
9. Sharma, A., Paliwal, K.K.: Linear discriminant analysis for the small sample size problem: an overview. International Journal of Machine Learning and Cybernetics **6**(3), 443–454 (June 2015)
10. Tao, D., Li, X., Wu, X., Maybank, S.J.: Geometric mean for subspace selection. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(2), 260–274 (2008)
11. Wang, H., Yan, S., Xu, D., Tang, X., Huang, T.: Trace ratio vs. ratio trace for dimensionality reduction. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2007)
12. Yao, C., Cheng, G.: Approximative bayes optimality linear discriminant analysis for chinese handwriting character recognition. Neurocomputing **207**, 346–353 (2016)
13. Zhang, Y., Yeung, D.Y.: Worst-case linear discriminant analysis. In: Advances in Neural Information Processing Systems. pp. 2568–2576 (2010)