Finding discriminant axes from multiple viewpoints

Jun FUJIKI
Fukuoka University
8-19-1, Nanakuma, Jonan-ku, Fukuoka 814-0180, JAPAN
fujiki@fukuoka-u.ac.jp

Abstract

Fisher's linear discriminant analysis (FLDA) has been widely used due to its simple formulation and low computational costs. However, FLDA implicitly assumes that all the classes share the same covariance, and therefore it might fail when this assumption is not necessarily satisfied. Among various extensions to tackle this problem, we focus on Detailed Fisher's linear discriminant analysis (DFDA) that can greatly improve the classification performance which preserving a simple formulation and low computational costs. However, its formulation seems ad-hoc and has not been theoretically justified yet. This paper proposes a new variant of DFDA that can be mathematically justified and intuitively comprehensive. Our new formulation reveals the fundamental nature of DFDA (and our new formulation) that tries to capture class-wise feature distributions from multiple views. Preliminary experiments demonstrate a promising result of our new formulation.

1 Introduction

Fisher's linear discriminant analysis (FLDA) [5] has been widely used as a discriminative feature extractor in the fields of pattern recognition, computer vision and machine learning [2, 8] for a long time due to its simple formulation and low computational costs. However, FLDA has a disadvantage: it implicitly assumes that a distribution of each class should be Gaussian and all the classes share the same covariance matrix. FLDA works very well when this assumption is satisfied, however, most real-world datasets are not in the case.

A lot of extensions and modifications of FLDA have ever been proposed to overcome the problem, which can be roughly classified into two approaches.

The first approach is non-linear or piecewise linear extentions. Hastie et al. [9], Zhu et al. [16], and Gkalelis et al. [7] integrated cluster analysis into FLDA to fit multi-peak feature distributions. Baudat [1] and Sierra [14] introduced non-linear formulations to deal with complex feature distributions. The first approach is very popular and yields outstanding performances against FLDA. However, it requires high computational costs, which would eliminate one of the strengths of FLDA. Furthermore, this approach often encounters some difficulty in model selection, such as the number of peaks and the type of transformations.

The second approach is the introduction of metrics between probabilistic distributions into the computation of between-class scatter matrices, instead of a simple Euclidean norm. Kullback-Leibler divergence and Chernoff distance [3, 11] has been tried for this purpose. One major problem of this approach lies on the asymmetric structure of metrics, which leads to inconsistent formulations of the entire method.

Sakano et al. [13] recently proposed yet another extension of FLDA called Detailed Fisher's linear discriminant analysis (DFDA). The main idea of DFDA is a combination of FLDA and the class description of Class Featuring Information Compression (CLAFIC) [12, 10]. Inspired by CLAFIC, DFDA injects covariance information of every class into a between-class scatter matrix by utilizing eigenvectors of class-specific auto-correlation matrices. These eigenvectors specify a feature subspace, and therefore have a potential to reveal a detailed covariance structure of every class. DFDA integrated them into the original FLDA by simply adding between-class scatter matrices derived from feature vectors and eigenvectors of class-wise auto-correlation matrices. The formulation of DFDA consists of simple matrix operations so that DFDA preserves the main advantages of FLDA, namely the simple formulation and low-computational costs. Despite of those merits, DFDA has a crucial drawback: No theoretical backgrounds and justifications. The way of the extension in DFDA seems ad-hoc, and the reason why DFDA achieves remarkable improvements has still been unclear.

This paper proposes a new variant of DFDA that can be mathematically comprehensive. Through the discussion in this paper, we clarify the fundamental nature of DFDA and our new formulation that tries to capture class-wise feature distributions from multiple viewpoints, one from the mean of all the features, the other from the origin. Although between-class scatters obtained from eigenvectors of class-wise auto-correlations would not be optimal from the standpoint of discriminant analysis, they would have a great potential to discriminative training from the analogy to CLAFIC.

The rest of the paper is organized as follows: Section 2 reviews the classical FLDA, and clarify its fundamental problems. Section 3 describes DFDA as an extension of FLDA, which integrates class-wise feature distributions. Section 4 proposes a new variant of DFDA by reformulating DFDA, which enables us to deeply understand the nature of DFDA. Section 5 reports preliminary experimental results with standard benchmark datasets. Finally Section 6 concludes the paper and poses some future work.

2 Fisher's linear discriminant analysis

This section reviews the classical FLDA. Briefly speaking, FLDA tries to maximize between-class distances and to minimize within-class distances. Especially, it finds a set of bases most discriminative for classifying features labeled with one of the C classes.

Let $\boldsymbol{x}_1^{(c)}, \ldots, \boldsymbol{x}_{n_c}^{(c)}$ be a set of D-dimensional sam-

ples in class c ($c = 1, \ldots, C$), where n_c is the number of samples assigned to the class c. Let $\boldsymbol{\mu}^{(c)}$ be the mean vector of samples assigned in the class c. Here, we use the notations $\boldsymbol{X}^{(c)} = (\boldsymbol{x}_1^{(c)} \quad \cdots \quad \boldsymbol{x}_{n_c}^{(c)})$ to represent the data matrix with class c, and $\boldsymbol{X} = \left(\boldsymbol{X}^{(1)} \quad \cdots \quad \boldsymbol{X}^{(C)}\right)$ to represent the data matrix composed of all the features.

The between-class distances can be characterized by the following between-class scatter matrix:

$$\begin{aligned}\boldsymbol{\Sigma}_B &= \frac{1}{n}\sum_{c=1}^{C} n_c(\boldsymbol{\mu}^{(c)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(c)} - \boldsymbol{\mu})^\top \\ &= \mathtt{var}[\boldsymbol{X}] - \frac{1}{n}\sum_{c=1}^{C} n_c \mathtt{var}[\boldsymbol{X}^{(c)}] \end{aligned} \quad (1)$$

where n is the total number of data, $\boldsymbol{\mu}$ is the mean vector of all samples, and $\mathtt{var}[\boldsymbol{A}]$ is the scatter matrix of columns of $\boldsymbol{A}$. The within-class distances can also be characterized by the following within-class scatter matrix:

$$\boldsymbol{\Sigma}_W = \frac{1}{n}\sum_{c=1}^{C} n_c \boldsymbol{\Sigma}^{(c)} = \frac{1}{n}\sum_{c=1}^{C} n_c \mathtt{var}[\boldsymbol{X}^{(c)}] \quad (2)$$

where $\boldsymbol{\Sigma}^{(c)} = \mathtt{var}[\boldsymbol{X}^{(c)}]$ is the scatter matrix of the class c. FLDA can be easily solved by the generalized eigenvalue problem $\boldsymbol{\Sigma}_B \boldsymbol{a} = \lambda \boldsymbol{\Sigma}_W \boldsymbol{a}$ where $\boldsymbol{a}$ is an eigenvector and λ is an eigenvalue obtained from the generalized eigenvalue problem. The eigenvectors obtained from the generalized eigenvalue problem correspond to the most discriminative axes for a given dataset $\boldsymbol{X}$ from the standpoint of FLDA.

The number of valid eigenvectors of the generalized eigenvalue problem is always less than C, since the between-class scatter matrix $\boldsymbol{\Sigma}_B$ is a sum of C one-rank matrices, and therefore its rank is less than C. This is one of critical disadvantages in FLDA. Also, it implicitly assumes that a distribution of each class is Gaussian and all the classes share the same covariance matrix. If this assumption is violated especially in high dimensional feature spaces, a lot of samples with difference class labels often tend to be distributed close with each other on the discriminant axes.

3 Detailed FLDA

This section reviews Detailed FLDA (DFDA) [13] that can be regarded as an extension of FLDA considering different covariance structures of each class.

As described in the previous section, FLDA only focuses on separating class mean vectors when calculating between-class distances. In other words, FLDA does not consider details of feature distributions such as covariance matrices of classes, which might have a potential for discrimination. To mitigate this problem, a straightforward extension of FLDA based on Kullback-Leibler divergence has been proposed by Decell et al [3]. However, it requires much larger computational costs than the original FLDA, which reduces the usefulness of FLDA. A much simpler extension with the Chernoff criterion has been proposed by Loog et al [11]. However, it does not scale to large class problems such as Chinese character classification, because it requires a number of pairwise binary classifications to deal with multi-class discrimination problems.

DFDA provides a much simpler and more effective solution to the above problem than the previous extensions. The main contribution of DFDA is the introduction of additional information inspired by CLAFIC [15]. In CLAFIC, a feature distribution of each class is represented by a subspace spanned by eigenvectors of the auto-correlation matrix of the class. Let $\boldsymbol{\psi}_k^{(c)}$ be the k-th eigenvector of the auto-correlation matrix $\boldsymbol{\Phi}^{(c)}$ of the c-th class, where

$$\boldsymbol{\Phi}^{(c)} = \frac{1}{n_c}\sum_{i=1}^{n_c} \boldsymbol{x}_i^{(c)} \boldsymbol{x}_i^{(c)\top}.$$

A subspace spanned by the eigenvectors $\boldsymbol{\psi}_k^{(c)}$ ($k = 1, 2, \cdots$) contains rich information about the feature distribution of the class c. This information is essentially different from the one contained in the within-class scatter matrix $\boldsymbol{\Sigma}_W$. To this end, DFDA incorporates a new criterion for evaluating disparity among classwise feature distributions into FLDA, with the use of the eigenvectors $\boldsymbol{\psi}_k^{(c)}$.

A newly introduced criterion of DFDA is defined as

$$\boldsymbol{\Sigma}_{B2} = \sum_{\substack{i,j=1 \\ i \neq j}}^{C} \sum_{k=1}^{d_u}\sum_{l=1}^{d_u} (\boldsymbol{\psi}_k^{(i)} - \boldsymbol{\psi}_l^{(j)})(\boldsymbol{\psi}_k^{(i)} - \boldsymbol{\psi}_l^{(j)})^\top, \quad (3)$$

where d_u is the number of eigenvectors of the auto-correlation matrix $\boldsymbol{\Phi}^{(c)}$ given in advance. Incorporating the above new criterion into FLDA, we can formulate DFDA with the generalized eigenvalue problem $(\boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_{B2})\boldsymbol{a} = \lambda \boldsymbol{\Sigma}_W \boldsymbol{a}$, namely the between-class matrix $\boldsymbol{\Sigma}_B$ is replaced by $\boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_{B2}$.

4 Proposed method and its analysis

This section describes our proposed method that is a simplified variant of DFDA to intuitively understand its nature.

First, let us introduce the following identity

$$\sum_{i=1}^{m}\sum_{j=1}^{m} (\boldsymbol{a}_i - \boldsymbol{a}_j)(\boldsymbol{a}_i - \boldsymbol{a}_j)^\top = 2m^2 \mathtt{var}[\boldsymbol{A}]$$

where $\boldsymbol{a}_i$ is the i-th column vector of $\boldsymbol{A}$.

This identity readily yields the following equations:

$$\begin{aligned}\boldsymbol{\Gamma}_T &= \sum_{i,j=1}^{C}\sum_{k=1}^{d_u}\sum_{l=1}^{d_u} (\boldsymbol{\psi}_k^{(i)} - \boldsymbol{\psi}_l^{(j)})(\boldsymbol{\psi}_k^{(i)} - \boldsymbol{\psi}_l^{(j)})^\top \\ &= 2(Cd_u)^2 \mathtt{var}[\boldsymbol{\Psi}], \end{aligned} \quad (4)$$

$$\begin{aligned}\boldsymbol{\Gamma}_W &= \sum_{c=1}^{C}\sum_{k=1}^{d_u}\sum_{l=1}^{d_u} (\boldsymbol{\psi}_k^{(c)} - \boldsymbol{\psi}_l^{(c)})(\boldsymbol{\psi}_k^{(c)} - \boldsymbol{\psi}_l^{(c)})^\top \\ &= \sum_{c=1}^{C} 2d_u^2 \mathtt{var}[\boldsymbol{\Psi}^{(c)}], \end{aligned} \quad (5)$$

$$\boldsymbol{\Psi}^{(c)} = (\boldsymbol{\psi}_1^{(c)} \ \ldots \ \boldsymbol{\psi}_{d_u}^{(c)}), \quad \boldsymbol{\Psi} = (\boldsymbol{\Psi}^{(1)} \ \ldots \ \boldsymbol{\Psi}^{(C)}).$$

This indicates that the criterion $\mathbf{\Sigma}_{B2}$ introduced in DFDA (cf. Eq. (3)) can be rewritten as

$$\begin{aligned}\mathbf{\Sigma}_{B2} &= \mathbf{\Gamma}_T - \mathbf{\Gamma}_W \\ &= 2C^2 d_u^2 \left\{ \mathtt{var}[\mathbf{\Psi}] - \frac{1}{C^2} \sum_{c=1}^{C} \mathtt{var}[\mathbf{\Psi}^{(c)}] \right\}. \quad (6)\end{aligned}$$

The above equation helps us to understand some mathematical background of DFDA. We can see from the equation that the new quantity $\mathbf{\Sigma}_{B2}$ introduced in DFDA exploits the sum of within-class scatters $\mathtt{var}[\mathbf{\Psi}^{(c)}]$ and whole the scatter $\mathtt{var}[\mathbf{\Psi}]$. Although this operation seems similar to the one of calculating the between-class scatter in FLDA, we can find two major differences between them: (1) In DFDA, all the scatters are calculated from eigenvectors of auto-correlation matrices, not from features, and (2) within-class scatters seem under-evaluated in DFDA, which can be seen by comparing Eq. (1) with Eq. (6). If we want to keep a balance between whole the scatter $\mathtt{var}[\mathbf{\Psi}]$ and class-wise scatters $\mathtt{var}[\mathbf{\Psi}^{(c)}]$, the between-class scatter $\mathbf{\Xi}_B$ and within-class scatter $\mathbf{\Xi}_W$ of eigenvectors of auto-correlation matrices should be defined as follows:

$$\mathbf{\Xi}_B = \mathtt{var}[\mathbf{\Psi}] - \frac{1}{C} \sum_{c=1}^{C} \mathtt{var}[\mathbf{\Psi}^{(c)}],$$

$$\mathbf{\Xi}_W = \frac{1}{C} \sum_{c=1}^{C} \mathtt{var}[\mathbf{\Psi}^{(c)}].$$

In general, eigenvectors of auto-correlation matrices has been already normalized, while features are not necessarily normalized. This implies that two types of within-class scatters ($\mathbf{\Sigma}_W$ and $\mathbf{\Xi}_W$) and between-class scatters ($\mathbf{\Sigma}_B$ and $\mathbf{\Xi}_B$) might have different ranges. To control the difference of ranges, we introduce a weight parameter α, resulting in the following within-class and between-class scatter matrices:

$$\mathbf{\Sigma}_{B_{\text{new}}} = \mathbf{\Sigma}_B + \alpha \mathbf{\Xi}_B, \quad \mathbf{\Sigma}_{W_{\text{new}}} = \mathbf{\Sigma}_W + \alpha \mathbf{\Xi}_W.$$

A new formulation can be obtained by the following generalized eigenvalue problem:

$$(\mathbf{\Sigma}_B + \alpha \mathbf{\Xi}_B)\boldsymbol{a} = \lambda(\mathbf{\Sigma}_W + \alpha \mathbf{\Xi}_W)\boldsymbol{a}, \quad (7)$$

which we call Revised DFDA (RDFDA).

From the formulation of RDFDA, we can see that the auto-correlation terms $\mathbf{\Xi}_B$ and $\mathbf{\Xi}_W$ weighted by α can be regarded as regularization terms to avoid over-compression of the feature space. A feature space obtained by FLDA for C-class classification is generally over-compressed with dimension less than C, which will lead to poor classification accuracy. This rank deficiency can be avoided if the rank of the auto-correlation term $\mathbf{\Xi}_B$ is sufficiently large, which might yield improvement in classification accuracy.

Our main claim in this paper is that observing features from multiple viewpoints is quite significant. The new formulation Eq.(7) consists of scatter terms ($\mathbf{\Sigma}_B$ and $\mathbf{\Sigma}_W$) and auto-correlation terms ($\mathbf{\Xi}_B$ and $\mathbf{\Xi}_W$). Here, we note that a scatter matrix can be regarded as an auto-correlation matrix when the origin of the coordinate system is located at the center of mass of features. From this aspect, RDFDA tries to capture class-wise feature distributions from two viewpoints, one from the center of mass of every class and the other is the origin of the coordinate system. The viewpoint from the center of mass is usual in various types of discriminant analyses, while the viewpoint from the origin of coordinate system has not been tried so far in any types of discriminant analyses, except DFDA. Observing features from the center of mass describes their relative positions, while observing features from the origin of the coordinate system reveals their absolute positions, which avoids over-compression of the features space.

The weighting parameter α takes an important role to improve the classification accuracy, since it controls the balance of two viewpoints as well as the degree of regularization. However in this paper, we simply set $\alpha = 1$ to check the effectiveness of our new formulation.

5 Experiments

In this section, experimental evaluations and demonstrates the effectiveness of DFDA and RDFDA is presented. In the experiments, a few datasets selected from UCI machine learning repository[1] is emploied.

The datasets are selected based on the two conditions as follows: (1) The number of classes, C, is smaller than dimension of the feature vectors, D. (2) The number of samples in each class, n_c, is larger than the dimension of the feature vectors, D.

After discriminant space is computed, data are classified by one nearest neighbor method.

Table 1 summarizes all the experimental results that include classification accuracy and the number of eigenvectors of class-wise auto-correlation matrices, d_u. In Table 1, the classification accuracy of FLDA is better than that in Sakano et al.[13] because of by selecting parameters and compress dimension d_u carefully in check experiments of Sakano et al.[13].

As shown in the table, the classification accuracies of RDFDA and DFDA are superior to FLDA for all the datasets. The results indicate that the regularization terms of RDFDA and DFDA were surprisingly effective.

The experimental results also show that RDFDA marked comparable performance with DFDA, and the superiority depends on dataset. However, the optimal number of eigenvectors taken from class-wise auto-correlation matrices for RDFDA was almost the same as or much smaller than that of DFDA, which indicates a potential of our new method RDFDA.

6 Conclusion

This paper proposes a new variant of FLDA called Revised DFDA (RDFDA). As shown in the name, our new method is a minor update of DFDA, however this update provides some theoretical justifications for RDFDA and DFDA. Our main claim in this paper is that observing features from multiple viewpoints is quite significant. FLDA only observes data from a single specific viewpoint, meaning the center of mass. Meanwhile, (R)DFDA observes data from two different viewpoints, one is the center of mass, and the other is

[1] `http://archive.ics.uci.edu/ml`

Table 1. Evaluation of proposed method on UCI MLR data

Data	D	C	♯ of training/test samples	FLDA	DFDA(d_u)	RDFDA(d_u)
Breast cancer	28	2	400/ 369	78.0%	**89.4%**(13)	87.8%(19)
magic	10	2	400/18820	55.8%	70.9%(4)	**75.3%**(5)
wine	13	3	180/ 118	40.0%	**91.5%**(6)	90.7%(6)
spambase	8	2	400/ 4401	55.8%	**83.8%**(28)	81.8%(29)
image segmentation	19	7	1470/ 2100	48.9%	88.4%(18)	**88.8%**(9)
ionosphere	34	2	200/ 251	56.6%	**89.6%**(24)	88.9%(17)
statlog(Landsat)	36	6	10800/ 4635	26.1%	**76.4%**(35)	73.5%(34)
statlog(Shuttle)	9	7	43500/14500	91.4%	99.8%(9)	**99.9%**(3)
statlog (vehicle)	18	4	1600/ 446	37.2%	70.4%(15)	**70.9%**(17)
madelon	500	2	4000/ 600	54.2%	**77.7%**(477)	55.8%(456)
optdigits	64	10	38230/ 1797	45.4%	**98.2%**(56)	95.1%(38)
Cardiotocography	21	3	3000/ 1126	73.5%	87.9%(6)	**88.5%**(5)

the origin of the coordinate system. Observing features from other viewpoints would work as a kind of regularization, which avoids rank deficiency FLDA often encounters. The effectiveness of our new formulation was shown by experiments with several datasets in UCI machine learning repository. Our new formulation are still composed of only simple matrix operations, and therefore we can enjoy low computational costs and high classification accuracy as DFDA do so.

Promising future work includes selecting optimal weighting parameters, integration with other types of extensions of FLDA such as [7, 9, 11, 16] and extensions to canonical correlation analysis.

Acknowledgement

The author would like to express grateful to Dr. Hitoshi Sakano and Dr. Akisato Kimura, the members of NTT comunication science laboratory, for valuable discussions.

References

[1] G. Baudat and F. Anouar, "Generalized Discriminant Analysis Using a Kernel Approach," Neural Computation, vol. 12, no. 10, pp. 2385–2404, 2006.

[2] P. N. Belhumeur et al. "Eigenfaces vs Fisherfaces: Recognition using class specific linear projection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, pp. 711–720, 1997.

[3] H. P. Decell and S. M. Mayekar, "Feature Combinations and the Divergence Criterion," Computers and Math. with Applications, vol. 3, pp. 71–76, 1977.

[4] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis,* John Wiley and Sons, 1973.

[5] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eigenics*, vol. 7, pp. 179–188, 1936.

[6] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* 2nd eds. Academic Press, 1990.

[7] N. Gkalelis, V. Mezaris and I. Kompatsiaris, "Mixture subclass discriminant analysis," IEEE Signal Processing Letters vol. 18, no. 5, pp. 319–322, 2011.

[8] T. Hastie, A. Buja and R. Tibshirani, "Penalized Discriminant Analysis," The Annals of Statistics, vol. 23, no. 1, pp. 73–102, 1995.

[9] T. Hastie and R. Tibshirani, "Discriminant Analysis by Gaussian Mixture," J. Royal Society of Statistical. Soc. B., vol. 58, pp. 155-176, 1996.

[10] G. Lim and C. H. Park, "Semi-supervised Dimension Reduction Using Graph-Based Discriminant Analysis," In: 2009 Ninth IEEE International Conference on Computer and Information Technology, pp. 9–13, 2009.

[11] M. Loog, and R. P. W. Duin, "Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 6, pp. 732–739, 2004.

[12] H. Sakano, "A Brief History of the Subspace Methods," In: Koch, R., Huang, F. (eds.) ACCV Workshops 2010, Part II. LNCS, vol. 6469, pp. 434–435. Springer, Heidelberg, 2011.

[13] H. Sakano et al. "Extended Fisher Criterion Based on Auto-correlation Matrix Information," In Proceedings of Structural, Syntactic, and Statistical Pattern Recognition - Joint IAPR International Workshop, SSPR&SPR 2012, pp. 409-416, 2012.

[14] A. Sierra, "High-order Fishers discriminant analysis," Pattern Recognition, vol. 35, no. 6, pp. 1291–1302, 2002.

[15] S. Watanabe, P. F. Lambert, C. A. Kulikowski, J. L. Buxton and R. Walker, "Evaluation and selection of variables in pattern recognition," Comp. & Info. Sciences, vol. 2, pp. 91–122, 1967.

[16] M. Zhu, A. M. Martinez, "Subclass discriminant analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 8, pp. 1274–1286, 2006.