Extended Fisher Criterion Based on Auto-correlation Matrix Information

Hitoshi Sakano[1], Tsukasa Ohashi[2], Akisato Kimura[1], Hiroshi Sawada[1], and Katsuhiko Ishiguro[1]

[1] NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
sakano.hitoshi@lab.ntt.co.jp
[2] Graduate School of Engineering, Doshisha University 1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto 610-0394, Japan

Abstract. Fisher's linear discriminant analysis (FLDA) has been attracting many researchers and practitioners for several decades thanks to its ease of use and low computational cost. However, FLDA implicitly assumes that all the classes share the same covariance: which implies that FLDA might fail when this assumption is not necessarily satisfied. To overcome this problem, we propose a simple extension of FLDA that exploits a detailed covariance structure of every class by utilizing revealed by the class-wise auto-correlation matrices. The proposed method achieves remarkable improvements classification accuracy against FLDA while preserving two major strengths of FLDA: the ease of use and low computational costs. Experimental results with MNIST and other several data sets in UCI machine learning repository demonstrate the effectiveness of our method.

1 Introduction

This paper proposes a simple extension of Fisher's linear discriminant analysis (FLDA) that exploits a detailed covariance structure of every class. The major advantage of our method lies on its ease of use and low computational cost. This makes our method more useful for practitioners.

FLDA is widely used as a discriminative feature extractor, especially in the field of pattern recognition, computer vision and machine learning. Its application areas have a wide variety, which include character recognition and face recognition [3,4]. FLDA has been attracting a lot of researchers and practitioners for a long time thanks to its simple formulation and low computational costs. However, FLDA implicitly assumes that a distribution of each class should be Gaussian and all the classes share the same covariance matrix. When facing a classification problem with other circumstances, its classification performance might degrade drastically.

Many extensions of FLDA have been proposed to overcome this problem. They are roughly classified into two categories. The first category is (1) non-linear or

G.L. Gimel' farb et al. (Eds.): SSPR & SPR 2012, LNCS 7626, pp. 409–416, 2012.

piecewise linear extensions of FLDA. Hastie, et al.[8],Zhu, et al.[9],Gkalelis, et al.[10] employed cluster analysis to fit multi-peak feature distributions. Baudat[6] and Sierra[7] studied non-linear transformation extension to represent complex feature distributions. This approach is very popular, however, it requires high computational cost that may eliminate one of the strengths of FLDA. Further, this approach may incur model selection difficulties such as the number of peaks and the type of transformations. Another approach is: (2) incorporating between-distribution metrics such as Kullback-Leibler divergence or Chernoff distance into the computation of between-class scatter matrices [13,12], instead of simple Euclidean norms. One major problem of this approach lies on the asymmetric structure of metrics, which leads to inconsistent formulations of the entire method. In other words, the second approach is attractive if we can avoid this problem.

Based on the above observations, this paper proposes yet another extension of FLDA along with the second approach. The main problem is how to inject covariance information of every class into between-class scatter matrix. Inspired by the class description of Class Featuring Information Compression (CLAFIC) [14,11], we describe this covariance information as a subspace spanned by eigenvectors of a class-specific auto-correlation matrix. Thus, we can acquire rich information of class-wise feature distributions by simply concatenating the subspace induced from auto-correlation matrix to the subspace obtained from the original FLDA. Our proposed formulation consists of simple matrix operations only, and the algorithm is still easy to use and enjoy low-computational cost. Further it is easy to extend the formulation to multi-class categorization problems.

The rest of the paper is organized as follows. Section 2 reviews the classical FLDA and clarify its fundamental problems. Section 3 describes our new criterion function for FLDA based on the description of class-wise feature distribution. Section 4 demonstrates the effectiveness of the proposed method through some experimental evaluations with standard benchmark datasets. Finally Section 5 concludes this paper and poses some future work.

2 Fisher's Discriminant Analysis and Its Problems

This section reviews the classical FLDA and clarifies its fundamental problems.

Let $X_c = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_c}\}$ be a set of D-dimensional samples in class c, where n_c is the number of samples assigned to the class c. To find the most discriminative basis for C-class classification problem, FLDA maximizes between-class distances represented as the following between-class scatter matrix:

$$\Sigma_B = \frac{1}{C}\sum_{c=1}^{C}(\boldsymbol{\mu} - \boldsymbol{\mu}_c)(\boldsymbol{\mu} - \boldsymbol{\mu}_c)^\top, \tag{1}$$

and minimizes within-class distances represented as the following within-class scatter matrix:

$$\Sigma_W = \frac{1}{C}\sum_{c=1}^{C}\Sigma_c, \tag{2}$$

where $\boldsymbol{\mu}$ is the mean vector of all samples, $\boldsymbol{\mu}_c$ is the mean vector of samples assigned in the class c, and Σ_c is the scatter matrix of the class c:

$$\Sigma_c = \frac{1}{n_c} \sum_{i=1}^{n_c} (\boldsymbol{\mu}_c - \boldsymbol{x}_{ci})(\boldsymbol{\mu}_c - \boldsymbol{x}_{ci})^\top. \tag{3}$$

The problem is easily solvable as the following generalized eigenvalue problem:

$$\Sigma_B \boldsymbol{a} = \Sigma_W \boldsymbol{a} \lambda, \tag{4}$$

where $\boldsymbol{a}$ is an eigenvector and λ is an eigenvalue obtained from the above generalized eigenvalue problem. Eigenvectors correspond to the most distinctive axes (projections) to the given dataset.

The number of valid eigenvectors of the above generalized eigenvalue problem should be less than C, since the number of class means is C and therefore the maximum rank of the between-class scatter matrix is $C - 1$ if $D > C - 1$. When dealing with 2-class classification problems, only one dimensional subspace is available. If common covariance assumption is violated in high dimensional feature space, since most of samples are distributed out of discriminant axis true discriminant plane may be close to discriminant axis. This dimensionality limitation is the fundamental problem of FLDA.

3 DFDA: Describing Covariance Structure of Classes in FLDA

In this section we propose a new FLDA criterion that reflects unique covariance structures of each class. Our observation is that one reason of the dimensionality limitation of FLDA is that FLDA only focuses on separating class mean vectors. In other words, FLDA does not consider the difference of covariance matrices of classes, or information about sample distributions of classes, which might have certain discriminative power to the classification problem. From this point of view, a straightforward extension of FLDA has been proposed in [12] that is based on Kullback-Leibler divergence. However, its optimization procedure is much complex than the original FLDA and it weakens the usefulness of FLDA. A more simpler extension, which is based on the Chernoff criterion, has been proposed in [13]. However, this method does not scale to large class problems such as Chinese character classification because this model requires pairwise classification procedure for multiclass problems.

Our proposed method is inspired CLAFIC[14]: representing sample distribution information of classes as subspaces. $\boldsymbol{\psi}_{ck}$ denotes the k-th eigenvector of the c-th class auto-correlation matrix Γ_c:

$$\Gamma_c = \frac{1}{n_c} \sum_{\boldsymbol{x} \in \omega_c} \boldsymbol{x}\boldsymbol{x}^\top. \tag{5}$$

$\boldsymbol{\psi}_{ck}$ is not an eigenvector of the covariance matrix, but obviously has information about the distribution of samples of the class c (because this is an eigenvector of auto-correlation). Our key idea is to use $\boldsymbol{\psi}_{ck}$ to compute dispersions between classes: intuitively, a classifier that separates $\boldsymbol{\psi}_{ck}$s from $\boldsymbol{\psi}$s of other classes is a good classifier because it segregates "shapes" of class distributions. Our criterion is described as maximization of

$$\Sigma_{B2} = \sum_{c} \sum_{d \neq c} \sum_{k=1}^{d_u} \sum_{l=1}^{d_u} (\psi_{ck} - \psi_{dl})(\psi_{ck} - \psi_{dl})^{\top}. \tag{6}$$

d_u denotes a number of eigenvectors of auto correlation matrix used for computation, which must be predefined by users. Based on this criterion, we define a new between scatter matrix as

$$\Sigma_{B_{new}} = \Sigma_B + \Sigma_{B2}. \tag{7}$$

If the rank of $\Sigma_{B_{new}}$, $r = \mathrm{rank}(\Sigma_{B_{new}})$ is greater than $C - 1$, then we can expect improvement of classification accuracy. We call this extension of FLDA as Detailed FLDA (DFDA), which maximizes $\Sigma_{B_{new}}$. A concept sketch of DFDA illustrate in Fig.1[1].

Figure 1(a) illustrates when classes share the same covariance as assumed implicitly in FLDA. In such a case, the FLDA provides optimal projections. However, FLDA is not optimal in the case (b) because classes have different covariances (distributions, or "shapes" of shaded regions). On the other hand, DFDA projection also tries to separate eigenvectors of class auto-correlation matrices ψ and there is no assumption of shared covariances among classes. Thus we can expect improvement of classification accuracy when covariances among classes are different.

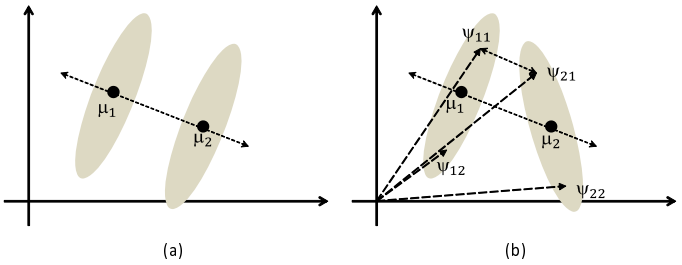

Fig. 1. Conceptual sketch of the proposed method. (a)conventional FLDA (b)proposed DFDA. The vectors ψ implicitly reveal the distributions (shapes) of the classes. The proposed DFDA tries to take the class distribution away from the others by incorporating ψ into the original FLDA.

[1] Though this figure is not truly correct, we expect the figure helps readers to understand the concept of the proposed method.

4 Experiments

In this section, we present the experimental evaluations. As the first experiment, we employ MNIST handwritten digit database [2]. The MNIST dataset is known as a standard benchmark data for statistical pattern recognition and machine learning researches. We would like to understand and present behaviors of proposed DFDA by this experiment. Second experiment employs a few datasets taken from UCI machine learning repository. A goal of the second experiment is to confirm effectiveness of the proposed DFDA over various datasets[3].

4.1 Experiments with MNIST Dataset

MNIST database consists of 10 digits (=classes, $C = 10$) handwritten character images. Images are 28×28 real valued matrices. We used vectorized matrices as feature vectors, thus feature space dimensionality is $D = 784$.

One of the characteristics of MNIST dataset is "FLDA hard."[15] So far many researchers have evaluated various classifiers and feature extractors with MNIST dataset. However the accuracy score is relatively low if we employ FLDA: for example classification accuracy of 1-NN classifier in original pixel value space is 97.3% that in FLDA space is 90.5%. We guess this difficulty is caused by dimension limitation of FLDA.

Since we would like to understand behaviors of FLDA and DFDA as feature extractors, we employ a 1-NN classifier in reduced spaces induced by FLDA or DFDA. We used PRTools 4.3[4] and a Matlab implementation of proposed method. We used 60000 samples as training data and 10000 samples as test data. We tested several values of d_u, the number of used eigenvectors ψ of auto-correlation matrices.

The results are shown in Table 1, Fig. 2, and Fig. 3.

Table 1. Classification accuracy of FLDA and DFDA

Method	d_u	r	Classification Accuracy (%)
FLDA	0	9	90.5
DFDA	50	51	94.3
DFDA	100	77	95.2
DFDA	200	120	95.3
DFDA	400	229	93.5

From the Table 1, it is obvious that the rank of augmented between-class scatter matrix $\Sigma_{B_{new}}$, r, grows as the number of used ψ, d_u, increases. This indicates that the information from auto-correlation matrices actually augments the information for class separations. Figure 2 illustrates the evolution of classification

[2] `http://yann.lecun.com/exdb/mnist`
[3] `http://archive.ics.uci.edu/ml`
[4] `http://www.prtools.org`

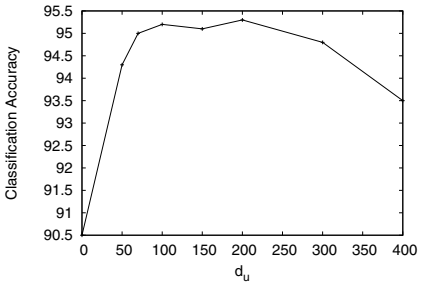

Fig. 2. Classification accuracy of proposed method. The horizontal axis denotes d_u, the number of used eigenvectors ψ of auto-correlation matrices. The vertical axis denotes the classification accuracy.

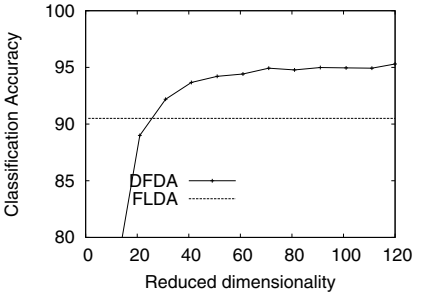

Fig. 3. Comparison of classification accuracy between FLDA and DFDA on MNIST dataset ($d_u = 200$). The horizontal axis denotes the dimensionality of the extracted features.

accuracies against d_u, the number of used eigenvectors ψ of auto-correlation matrices in DFDA. The classification accuracy score hits the highest at $d_u = 200$. This result indicates that there is an balancing point in adding ψs, possibly because of the ranks of auto-correlations matrices: these auto-correlations matrices may not be full-rank, too.

Finally, Fig.3 shows a comparison of classification accuracy of the original FLDA and the proposed DFDA ($d_u = 200$). The horizontal axis denotes the dimensionality of the extracted features. In other words, the number of eigenvectors (projections) obtained by DFDA. Note that the dimensionality of extracted features by FLDA is $9 = C - 1$. If we employ very few number of eigenvectors (less than 20), the DFDA performs poor, even worse than FLDA. However, as the number of eigenvectors increases, the performance of DFDA outperforms the FLDA, and saturates around 60 dimensions.

4.2 Experiments with Dataset from UCI Machine Learning Repository

To confirm effectiveness of DFDA, we evaluated the proposed DFDA with several datasets from UCI machine learning repository. The selection of datasets is based on the following conditions:

- The number of classes, C, is smaller than dimensionality of the feature vectors, D.
- The number of samples in each class, n_c, is larger than the dimensionality of the feature vectors, D.

Table 2 summarizes the computed classification accuracies. As evident from the table, DFDA surprisingly performs better than the FLDA in all the datasets.

Table 2. Evaluation of proposed method on UCI MLR data

Data	D	C	# of training samples ($n_c \times C$)	r	FLDA	DFDA
Breast cancer	9	2	200	8	78.0%	79.9%
magic	10	2	200	8	55.8%	69.6%
wine	13	3	60	13	40.0%	72.9%
spambase	8	2	200	8	55.8%	62.7%
image segmentation	19	7	210	19	48.9%	86.0%
ionosphere	34	2	100	8	56.6%	75.7%
statlog(Landsat)	36	6	1800	36	26.1%	75.3%
statlog(Shuttle)	9	7	43500	9	91.4%	99.7%
statlog (vehicle)	18	4	400	18	37.2%	69.7%
madelon	500	2	2000	500	54.2%	60.7%
optdigits	64	10	3823	64	45.4%	97.9%
Cardiotocography	21	3	1000	21	73.5%	78.3%

5 Conclusion

This paper proposed an extension of Fisher's Linear Discriminant Analysis (FLDA) by injecting inherent differences of distributions among classes. The proposed method exploited the auto-correlation matrix of each class samples inspired by CLAFIC. The proposed Detailed Fisher Discriminant Analysis (DFDA) integrates the subspace spanned by eigenvectors obtained from the auto-correlation matrix into the between-class scatter matrix of FLDA. Experimental evaluations with MNIST dataset and several dataset in UCI machine learning repository demonstrated the effectiveness of proposed method. Our proposed method is composed of only simple matrix operations, and therefore it can be naturally applied to multi-class categorization. The method might provide some new direction of FLDA.

The major weakness of the proposed method is its theoretical foundations. Since our idea is intuitively sound, we need more theoretical justification for this extension. It is also interesting to compare the proposed method with other extensions of FLDA such as [13,9,10,8]. Finally, some extensions of canonical correlation analysis can be achieved in a similar way, which might be fruitful for many applications.

References

1. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of Eigenics 7, 179–188 (1936)
2. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. John Willey and Sons (1973)
3. Belhumeur, P.N., et al.: Eigenfaces vs Fisherfaces: Recognition using class specific linear projection. IEEE Transaction of Pattern analysis and Machine PAMI 19, 711–720 (1997)

4. Hastie, T., Buja, A., Tibshirani, R.: Panalized Discriminant Analysis. The Annals of Statistics 23(1), 73–102 (1995)
5. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press (1990)
6. Baudat, G., Anouar, F.: Generalized Discriminant Analysis Using a Kernel Approach. Neural Computation 12(10), 2385–2404 (2006)
7. Sierra, A.: High-order Fishers discriminant analysis. Pattern Recognition 35(6), 1291–1302 (2002)
8. Hastie, T., Tibshirani, R.: Discriminant Analysis by Gaussian Mixture. J. Royal Society of Statistical. Soc. B. 58, 155–176 (1996)
9. Zhu, M., Martinez, A.M.: Subclass discriminant analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(8), 1274–1286 (2006)
10. Gkalelis, N., Mezaris, V., Kompatsiaris, I.: Mixture subclass discriminant analysis. IEEE Signal Processing Letters 18(5), 319–322 (2011)
11. Sakano, H.: 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)
12. Decell, H.P., Mayekar, S.M.: Feature Combinations and the Divergence Criterion. Computers and Math. with Applications 3, 71–76 (1977)
13. Loog, M., Duin, R.P.W.: Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(6), 732–739 (2004)
14. Watanabe, S., Lambert, P.F., Kulikowski, C.A., Buxton, J.L., Walker, R.: Evaluation and selection of variables in pattern recognition. Comp. & Info. Sciences 2, 91–122 (1967)
15. Lim, G., Park, C.H.: Semi-supervised Dimension Reduction Using Graph-Based Discriminant Analysis. In: 2009 Ninth IEEE International Conference on Computer and Information Technology, pp. 9–13 (2009)