A Discriminative and Heteroscedastic Linear Feature Transformation for Multiclass Classification

Hung-Shin Lee[1,2], Hsin-Min Wang[1], Berlin Chen[3]
[1]*Institute of Information Science, Academia Sinica, Taiwan*
[2]*Dept. of Electrical Engineering, National Taiwan University, Taiwan*
[3]*Dept. of Computer Science & Information Engineering, National Taiwan Normal University*
E-mail: {hslee, whm}@iis.sinica.edu.tw, berlin@ntnu.edu.tw

Abstract

This paper presents a novel discriminative feature transformation, named full-rank generalized likelihood ratio discriminant analysis (fGLRDA), on the grounds of the likelihood ratio test (LRT). fGLRDA attempts to seek a feature space, which is linearly isomorphic to the original n-dimensional feature space and is characterized by a full-rank $(n \times n)$ *transformation matrix, under the assumption that all the class-discrimination information resides in a d-dimensional subspace* $(d < n)$*, through making the most confusing situation, described by the null hypothesis, as unlikely as possible to happen without the homoscedastic assumption on class distributions. Our experimental results demonstrate that fGLRDA can yield moderate performance improvements over other existing methods, such as linear discriminant analysis (LDA) for the speaker identification task.*

1. Introduction

For the purposes of better discrimination and less computational complexity, feature extraction by reducing the feature dimensionality is indispensable and crucial to the development of a pattern recognition system. It often aims to seek a linear transformation for projecting feature vectors from an original n-dimensional space to a d-dimensional subspace $(d < n)$, so that the resulting new features can possess good discriminatory power among classes. One of the most widely used methods is linear discriminant analysis (LDA), which can be thought of as a procedure that maximizes the average squared Mahalanobis distance between each class-mean pair in the projective subspace [1]. Apart from the above-mentioned geometric interpretation, Campbell has shown that the derivation of the LDA transformation is equivalent to finding the parameters of multivariate Gaussian models by means of maximum likelihood (ML) estimation, under the assumption that the whole class discrimination information resides in a d-dimensional subspace and that the within-class covariance matrices are equal for all classes (Fig. 1(a)) [2]. Afterwards, Kumar proposed heteroscedastic linear discriminant analysis (HLDA) to generalize LDA by dropping the homoscedastic assumption that all classes have equal within-class covariance matrices and maximizing the likelihood for these Gaussian models iteratively [3].

Accordingly, we can roughly summarize two common components in LDA and HLDA. First, the transformation matrix is derived by maximizing the likelihood of all samples in the projective subspace. Second, the whole information for class discrimination resides in the d-dimensional subspace, spanned by d column vectors of the transformation matrix. In other words, the rejected subspace does not possess any discriminatory power, where we can suppose that the distributions of all classes completely overlap.

This paper presents a novel discriminative feature transformation, named full-rank generalized likelihood ratio discriminant analysis (*f*GLRDA), stemming from the generic idea of the likelihood ratio test (LRT) and some parts of our previous work [8]. *f*GLRDA attempts to seek a feature space, which is linearly isomorphic to the original n-dimensional feature space and can be decomposed into a d-dimensional discriminatory subspace and an $(n-d)$-dimensional non-discriminatory subspace by making the most confusing situation, described by the null hypothesis, as unlikely as possible to happen without the homoscedastic assumption on underlying class distributions (Fig. 1(b)). The highlights of *f*GLRDA are summarized as follows:

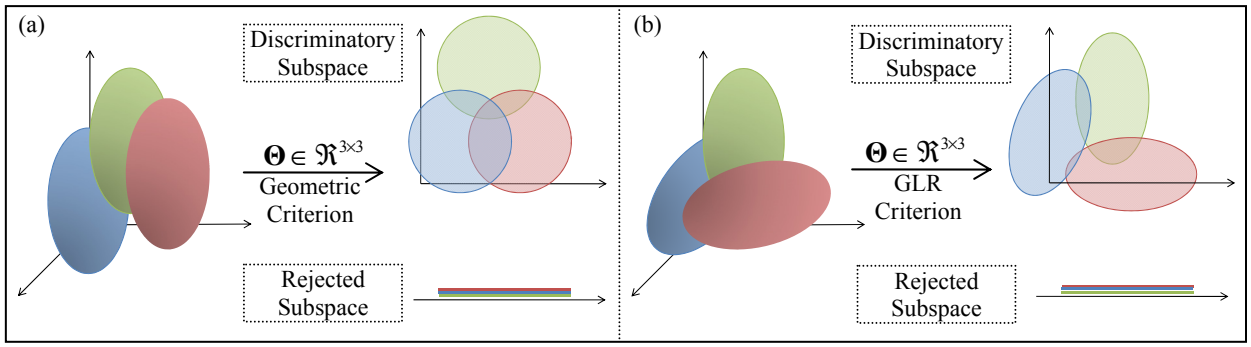

Figure 1. Illustrations of LDA (a) and *f*GLRDA (b) for three-class feature transformation.

1. Inheriting from LDA, *f*GLRDA guarantees the least discriminatory power in the rejected subspace.
2. As we shall see shortly, without the homoscedastic assumption, *f*GLRDA is amenable to pair with more elaborate classifiers like quadratic discriminant functions when compared to LDA.

2. The modified likelihood ratio test

2.1. Background

Conceptualized from statistical hypothesis testing [4], the likelihood ratio test (LRT) is a celebrated method of obtaining test statistics in situations where one wishes to test a null hypothesis H_0 against a completely general alternative hypothesis H_1. In this paper, H_0 generally represents a statistical fact that we would not like to accept. If Ω denotes the complete parameter space and ω denotes the parameter space restricted by the null hypothesis H_0, the LRT criterion for the null hypothesis H_0 against the alternative hypothesis H_1 is

$$LR = \frac{\max L_{\omega}}{\max L_{\Omega}} \qquad (1)$$

where L denotes the likelihood of the sampled data, and $\max L_S$ denotes the likelihood computed with the ML estimated parameter set S.

The logic behind the LRT criterion lies in that, if H_0 is apparently false with no extra confidence measure being considered, the ML condition might occur at a point in Ω other than ω, which means that $\max L_{\omega}$ will be far smaller than $\max L_{\Omega}$ ideally [4].

2.2. Problem formulation

In this paper, we, however, do not intend to strictly follow the LRT procedure for reducing the dimensionality of feature vectors. More specifically, we do not set the goal at testing whether the null hypothesis is true or false, but instead, at seeking a projected space, where the (most confusing) null hypothesis is as unlikely as possible to be true. To get to this point, we design the following statistical hypotheses:

$$\begin{cases} H_0\text{: The class populations are the same.} \\ H_1\text{: The class populations are different.} \end{cases}$$

The transformed space spanned by the column vectors of the nonsingular transformation matrix $\Theta \in \Re^{n\times n}$ must satisfy the condition that the likelihood of all feature vectors generated by the null hypothesis is as small as possible. In light of this, the objective function of full-rank generalized likelihood ratio discriminant analysis (*f*GLRDA) can be generically formulated by

$$J_{\text{fGLRDA}}(\Theta) = \frac{\max L_{\substack{\text{the parameter space that the} \\ \text{class populations are the same.}}}(\Theta)}{\max L_{\text{the complete paremeter space}}(\Theta)}. \qquad (2)$$

Finally, the transformation matrix Θ can be derived by minimizing $J_{\text{fGLRDA}}(\Theta)$.

3. Full-rank GLRDA

3.1. The model assumptions

Suppose the data is a collection of N independent labeled pairs $(\mathbf{x}_i, l_i)$, where $\mathbf{x}_i \in \Re^{n\times 1} (i \in \{1,...,N\})$ is a feature vector, and $l_i \in \{1,...,C\}$ is a class label. Each class $j \in \{1,...,C\}$ with the sample size n_j is modeled by a Gaussian distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$. The log-likelihood of the data in the transformed space is given by

$$\log L(\Theta) = (-Nd/2)\log(2\pi) + N\log|\Theta| - \sum_{j=1}^{C} \frac{n_j}{2} \begin{pmatrix} (\tilde{\mathbf{m}}_j - \tilde{\boldsymbol{\mu}}_j)^T \tilde{\boldsymbol{\Sigma}}_j^{-1} (\tilde{\mathbf{m}}_j - \tilde{\boldsymbol{\mu}}_j) + \\ \operatorname{tr}(\tilde{\boldsymbol{\Sigma}}_j^{-1} \tilde{\mathbf{S}}_j) + \log|\tilde{\boldsymbol{\Sigma}}_j| \end{pmatrix}, \qquad (3)$$

where $\mathbf{m}_j$ and $\mathbf{S}_j$ denote the sample mean vector and the sample covariance matrix of class j, and each variable with a tilde refers to the transformed version of the original variable. The term $|\Theta|$ in (3) comes from the Jacobian of the linear transformation Θ in accordance with the change of variables theorem.

Table 1. The MLE statistics of *f*GLRDA under various hypotheses.

Statistical Hypotheses	ML Estimates				(Relevant) Maximum Log-likelihood (not including the term $N\log\|\mathbf{\Theta}\|$)
	Mean Vectors		Covariance Matrices		
	Discriminatory	Rejected	Discriminatory	Rejected	
$H_0^{\text{hetero}}\begin{cases}\mathbf{\Sigma}_j:\text{unrestricted}\\ \mathbf{\mu}_j=\mathbf{\mu}\end{cases}$	$\mathbf{\Theta}_d^T\overline{\mathbf{w}}$	$\mathbf{\Theta}_{(n-d)}^T\overline{\mathbf{m}}$	$\mathbf{\Theta}_d^T\mathbf{W}_j\mathbf{\Theta}_d$	$\mathbf{\Theta}_{(n-d)}^T\mathbf{S}_T\mathbf{\Theta}_{(n-d)}$	$-\sum_{j=1}^{C}\frac{n_j}{2}\log(\|\mathbf{\Theta}_d^T\mathbf{W}_j\mathbf{\Theta}_d\|\,\|\mathbf{\Theta}_{(n-d)}^T\mathbf{S}_T\mathbf{\Theta}_{(n-d)}\|)$
$H_1^{\text{hetero}}\begin{cases}\mathbf{\Sigma}_j:\text{unrestricted}\\ \mathbf{\mu}_j:\text{unrestricted}\end{cases}$	$\mathbf{\Theta}_d^T\mathbf{m}_j$	$\mathbf{\Theta}_{(n-d)}^T\overline{\mathbf{m}}$	$\mathbf{\Theta}_d^T\mathbf{S}_j\mathbf{\Theta}_d$	$\mathbf{\Theta}_{(n-d)}^T\mathbf{S}_T\mathbf{\Theta}_{(n-d)}$	$-\sum_{j=1}^{C}\frac{n_j}{2}\log(\|\mathbf{\Theta}_d^T\mathbf{S}_j\mathbf{\Theta}_d\|\,\|\mathbf{\Theta}_{(n-d)}^T\mathbf{S}_T\mathbf{\Theta}_{(n-d)}\|)$

Moreover, we assume that only the first d components of $\tilde{\mathbf{x}}_i$ carry the class discrimination information. That is, let the first d columns of the full-rank linear transformation $\mathbf{\Theta}$ span the d-dimensional subspace, where the class mean vectors and the class covariance matrices are different. Therefore, the Gaussian parameters in the transformed space and the transformation matrix can be expressed by

$$\tilde{\mathbf{\mu}}_j=[\tilde{\mathbf{\mu}}_j^d,\tilde{\mathbf{\mu}}_0]^T=[\mathbf{\Theta}_d^T\mathbf{\mu}_j^d,\mathbf{\Theta}_{(n-d)}^T\mathbf{\mu}_0]^T, \quad (4)$$

$$\tilde{\mathbf{\Sigma}}_j=\begin{bmatrix}\mathbf{\Theta}_d^T\mathbf{\Sigma}_j^d\mathbf{\Theta}_d & 0\\ 0 & \mathbf{\Theta}_{(n-d)}^T\mathbf{\Sigma}_0^{(n-d)}\mathbf{\Theta}_{(n-d)}\end{bmatrix}, \quad (5)$$

where $\tilde{\mathbf{\Sigma}}_j^d$ and $\tilde{\mathbf{\Sigma}}_0^{(n-d)}$ are $d\times d$ and $(n-d)\times(n-d)$ matrices, respectively, and $\mathbf{\Theta}_d$ is composed of the first d columns of $\mathbf{\Theta}$ while $\mathbf{\Theta}_{(n-d)}$ the rest $(n-d)$ columns. Note that in the isomorphic space transformed by $\mathbf{\Theta}$, the first d-dimensional variables are totally uncorrelated with those in the remaining $(n-d)$-dimensional subspace.

3.2. The heteroscedastic case

In general, the parameter space, where the class populations are the same, can be characterized by the estimates of the class mean vectors. Therefore, in the heteroscedastic case that the covariance matrices of all classes are assumed to be different, the hypotheses of *f*GLRDA can be stated by

$$\begin{cases}H_0^{\text{hetero}}: \text{For class } j,\ \mathbf{\mu}_j=\mathbf{\mu} \text{ and } \mathbf{\Sigma}_j \text{ is unrestricted.}\\ H_1^{\text{hetero}}: \text{For class } j,\ \mathbf{\mu}_j \text{ and } \mathbf{\Sigma}_j \text{ are unrestricted.}\end{cases}$$

H_0^{hetero} describes an extreme situation that if it is true, the distributions of all class populations will become almost indistinguishable, resulting in less class-discrimination information offered by the parameter space. Therefore, the goal of *f*GLRDA is to find out the most appropriate projective subspace that makes the likelihood of the null hypothesis H_0^{hetero} as small as possible. The objective function of homoscedastic *f*GLRDA, which needs to be minimized, can be logarithmically expressed as

$$J(\mathbf{\Theta})=\max\log L_{H_0^{\text{hetero}}}(\mathbf{\Theta})-\max\log L_{H_1^{\text{hetero}}}(\mathbf{\Theta}). \quad (6)$$

The log-likelihood functions in (6) can be maximized with respect to their parameters under the constraints given by H_0^{hetero} and H_1^{hetero}. For the sake of simplicity, we can first derive the constrained estimates of the mean vectors and covariance matrices in terms of a fixed linear transformation $\mathbf{\Theta}$, rather than straightforwardly maximizing the log-likelihood functions. By differentiating the log-likelihood functions, which are defined by (3) and constrained by H_0^{hetero} and H_1^{hetero}, respectively, with respect to the corresponding parameters $\mathbf{\mu}_j$ and $\mathbf{\Sigma}_j$, and finding the point where the partial derivatives are zero, the ML estimates for each class can be derived. They are summarized in Table 1, where $\overline{\mathbf{m}}$ and $\mathbf{S}_T$ denote the global sample mean vector and the total scatter matrix of the data, respectively (cf. [5]). We also refer to $\overline{\mathbf{w}}$ as the weighted mean vector, which is computed by

$$\overline{\mathbf{w}}=\left(\sum_{j=1}^{C}n_j\mathbf{\Sigma}_j^{-1}\right)^{-1}\sum_{j=1}^{C}n_j\mathbf{\Sigma}_j^{-1}\mathbf{m}_j, \quad (7)$$

and $\mathbf{W}_j=(\mathbf{m}_j-\overline{\mathbf{w}})(\mathbf{m}_j-\overline{\mathbf{w}})^T+\mathbf{S}_j=\mathbf{B}_j+\mathbf{S}_j$. Note that in (7) $\overline{\mathbf{w}}$ contains an unknown term $\mathbf{\Sigma}_j$ that needs to be estimated. We can take $\mathbf{S}_j$ as a temporary estimate of $\mathbf{\Sigma}_j$ to form a sampled weighted mean for all classes. Substituting the maximized log-likelihoods of H_0^{hetero} and H_1^{hetero} in (6) with the values in Table 1, the objective function $J(\mathbf{\Theta})$ can be derived as

$$J(\mathbf{\Theta})=\sum_{j=1}^{C}-\frac{n_j}{2}\log(\|\mathbf{\Theta}_d^T\mathbf{W}_j\mathbf{\Theta}_d\|/\|\mathbf{\Theta}_d^T\mathbf{S}_j\mathbf{\Theta}_d\|). \quad (8)$$

Then, the derivative of $J(\mathbf{\Theta})$ is given by

$$\frac{\partial J(\mathbf{\Theta})}{\partial\mathbf{\Theta}}=-\sum_{j=1}^{C}n_j\frac{(-\mathbf{S}_j\mathbf{\Theta}\tilde{\mathbf{S}}_j^{-1}\tilde{\mathbf{B}}_j+\mathbf{B}_j\mathbf{\Theta})\tilde{\mathbf{S}}_j^{-1}}{1+\text{tr}(\tilde{\mathbf{S}}_j^{-1}\tilde{\mathbf{B}}_j)}, \quad (9)$$

Since $\partial J(\mathbf{\Theta})/\partial\mathbf{\Theta}=0$ has no analytical solution for the stationary points, we can use a gradient descent-based procedure for the minimization of $J(\mathbf{\Theta})$ [6].

4. Experiments and results

We evaluated the proposed *f*GLRDA on a series of speaker identification experiments conducted on the Japanese vowel dataset of the UCI-KDD archive [7]. The dataset contains discrete speech utterances of two Japanese vowels /a/ and /e/ pronounced by 9 male speakers. Each speaker has 54-118 utterances, and

there are 640 utterances in total. These utterances were represented by time-series recordings of 12-dimentional linear predictive coding (LPC) cepstral vectors with a stream length of 7-29. For each speaker, we used 30 time series to train his Gaussian discriminant function [5], and used the remaining 24-88 time series for testing.

In the training phase, each feature vector labeled to class i (i.e., uttered by speaker i) is thought of as an individual training instance. Based on the labeled data, we collect the corresponding statistics to derive the transformation matrices on top of the above mentioned methods, such that the classifiers can be generated by the transformed training vectors. However, in the test phase, given the Gaussian discriminant functions derived for each class in the training phase [5], each test utterance is first represented by its sample mean vector, which is taken as an input to the discriminant functions. The test utterance will be classified into class i that has the largest discriminant score.

Since the rank of the between-class scatter matrix is $C-1$, the maximum dimensionality of the discriminatory subspace generated by LDA is 8 ($C = 9$). However, *f*GLRDA can get around this limitation.

Figures 2 and 3 illustrate the plots of classification error rates as functions of the corresponding feature dimensions in the transformed subspace with linear and quadratic discriminant functions, respectively. It is clear that HLDA yields higher error rates, especially in the cases of lower dimensions. Since the maximum likelihood-based criterion of HLDA fails to account for class mean scatters, like the term $(\mathbf{m}_j - \overline{\mathbf{w}})(\mathbf{m}_j - \overline{\mathbf{w}})^T$, given by the null hypothesis H_0^{hetero}, which can be deemed to be a measurement of class separability in the discriminatory subspace, it does not necessarily generate discriminative effects. On the contrary, *f*GLRDA achieves the lowest error rates when the dimensionality is set to 3 or 4, which have relative error rate reductions of 22.63% and 51.17%, respectively, over LDA in association with the quadratic discriminant functions.

5. Conclusions

In this paper, we have presented a full-rank generalized likelihood ratio discriminant analysis (*f*GLRDA) approach for discriminative feature transformation on the basis of the likelihood ratio test. We argue that methods designed along this vein would be more feasible when being applied to a wide array of pattern recognition tasks. As part of future work, the parameter settings in the rejected subspace can be designed in a more elaborate way.

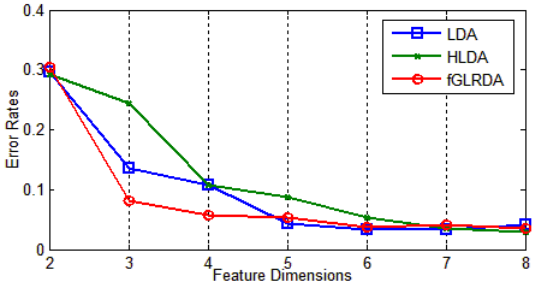

Figure 2. Plot of classification error rates versus feature dimensions with *linear* discriminant functions.

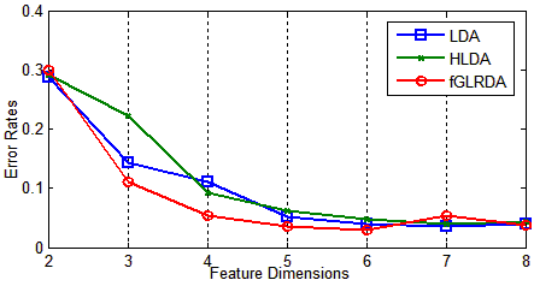

Figure 3. Plot of classification error rates versus feature dimensions with *quadratic* discriminant functions.

6. Acknowledgement

This work was supported in part by the National Science Council, Taiwan, under Grants: NSC96-2628-E-001-024-MY3, NSC96-2628-E-003-015-MY3, and NSC98-2631-S-003-002

7. References

[1] H.-S. Lee and B. Chen, "Empirical error rate minimization based linear discriminant analysis," in *Proc. ICASSP 2009*.

[2] N. A. Campbell, "Canonical variate analysis - a general model formulation," *Australian Journal of Statistics,* vol. 26, pp. 86-96, 1984.

[3] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication,* vol. 26, pp. 283-297, 1998.

[4] W. J. Krzanowski, *Principles of Multivariate Analysis: A User's Perspective*. Oxford University Press, 1988.

[5] R. O. Duda, *et al.*, *Pattern Classification*, 2nd ed., Wiley Press, 2001.

[6] R. Fletcher, *Practical Methods of Optimization*, Wiley Press, 2nd ed., 1987.

[7] M. Kudo, *et al.*, "Multidimensional curve classification using passing-through regions," *Pattern Recognition Letters*, vol. 20, no. 11, pp. 1103-1111, 1999.

[8] H.-S. Lee and B. Chen, "Generalized likelihood ratio discriminant analysis," in *Proc. IEEE ASRU 2009*.