



Rapid and brief communication

Uncorrelated heteroscedastic LDA based on the weighted pairwise Chernoff criterion

A.K. Qin^a, P.N. Suganthan^{a,*}, M. Loog^b^a*School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang Avenue, Block S2, Singapore 639798, Singapore*^b*Department of Innovation, Image Analysis Group, IT University of Copenhagen, Rued Langgards Vej 7, DK-2300 Copenhagen, Denmark*

Received 7 September 2004; accepted 20 September 2004

Abstract

We propose an uncorrelated heteroscedastic LDA (UHLDA) technique, which extends the uncorrelated LDA (ULDA) technique by integrating the weighted pairwise Chernoff criterion. The UHLDA can extract discriminatory information present in both the differences between per class means and the differences between per class covariance matrices. Meanwhile, the extracted feature components are statistically uncorrelated the maximum number of which exceeds the limitation of the ULDA. Experimental results demonstrate the promising performance of our proposed technique compared with the ULDA.

© 2004 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Uncorrelated linear discriminant analysis; Heteroscedastic; Weighted pairwise Chernoff criterion

1. Introduction

In statistical pattern recognition, linear dimensionality reduction (LDR) techniques are widely applied to reduce the complexity of the statistical model and often result in the improved classification accuracy in the transformed lower-dimensional space. Fisher's linear discriminant analysis (LDA) is one of the most popular supervised linear dimensionality reduction techniques, which tries to find an optimal set of discriminant vectors $\mathbf{W} = [\varphi_1, \dots, \varphi_d]$ by maximizing the Fisher criterion: $J_F(\mathbf{W}) = |\mathbf{W}^T \mathbf{S}_b \mathbf{W}| / |\mathbf{W}^T \mathbf{S}_w \mathbf{W}|$. Here, \mathbf{S}_b and \mathbf{S}_w are the between-class scatter matrix and average within-class scatter matrix of the training sample

group, respectively, which can be estimated as follows:

$$\begin{aligned} \mathbf{S}_b &= \sum_{i=1}^C P_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \\ &= \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T \quad \text{and} \\ \mathbf{S}_w &= \sum_{i=1}^C P_i \mathbf{S}_i, \end{aligned} \quad (1)$$

where C , P_i , \mathbf{m}_i , \mathbf{m} and \mathbf{S}_i represent the total number of pattern classes, a priori probability of pattern class ω_i , the mean vector of class ω_i , the mean vector of all training samples and the covariance matrix of class ω_i , respectively. The between-class scatter matrix \mathbf{S}_b can be expressed by both the original definition and its equivalent pairwise decomposition form [1].

* Corresponding author. Tel.: +00 65 6790 5404;

fax: +00 65 6792 0415.

E-mail addresses: qinkai@pmail.ntu.edu.sg (A.K. Qin),

epsugan@ntu.edu.sg (P.N. Suganthan), marco@isi.uu.nl

(M. Loog).

Uncorrelated features are usually desirable in pattern recognition tasks because an uncorrelated feature set is likely to contain more discriminatory information than a correlated one of the same dimension. Recently, Jin et al. [2] proposed the uncorrelated LDA technique (ULDA), which can obtain discriminant vectors by maximizing the Fisher criterion under the constraints that the extracted feature components are statistically uncorrelated, i.e. the derived discriminant vectors are subject to the \mathbf{S}_t -orthogonal constraints: $\varphi_i \mathbf{S}_t \varphi_j = 0, \forall i \neq j, i, j = 1, \dots, d$. Yang et al. [4] also demonstrated that ideal discriminant vectors should not only correspond to maximal Fisher criterion values but also correspond to minimal correlations between the extracted feature components. Therefore, the ULDA can yield a set of discriminant vectors with better discriminating power as shown experimentally in Refs. [2,4].

However, the ULDA technique still suffers from some deficiencies: firstly, it is incapable of dealing with heteroscedastic data in a proper way due to the implicit assumption that the covariance matrices for all the classes are equal. Hence, the derived discriminant vectors by the ULDA can merely attempt to separate the class means as much as possible while ignoring the discriminatory information present in the differences between the per class covariance matrices. This fact leads to the upper bound of the number of discriminant vectors extracted by ULDA to be limited to $C - 1$ as proven in Ref. [2]. Secondly, from the equivalent pairwise decomposition expression of the \mathbf{S}_b matrix, we can easily find that the class pair with large distance between them in the original feature space are overemphasized in the pairwise \mathbf{S}_b formula, which results in the obtained transformation attempting to preserve the distances of already well separated classes while causing larger overlap between pairs of classes that are not well separated in the original feature space. Consequently, the discriminant directions that may well separate the neighboring classes in the original feature space cannot be obtained by the ULDA if there are some classes far away and well separated from some other classes. In this paper, we propose an uncorrelated heteroscedastic LDA (UHLDA) technique based on the weighted pairwise Chernoff criterion, which can successfully solve the above problems.

2. Uncorrelated heteroscedastic LDA technique

2.1. ULDA technique

Suppose that \mathbf{S}_b and \mathbf{U} are positive semi-definite matrices and \mathbf{S}_w is a positive definite matrix. The first ULDA discriminant vector, denoted by φ_1 , is calculated as the eigenvector corresponding to the maximal eigenvalue of the eigenequation $\mathbf{S}_b \varphi = \lambda \mathbf{S}_w \varphi$. Suppose that i eigenvectors $\varphi_1, \varphi_2, \dots, \varphi_i, i \geq 1$, have been obtained. The $(i + 1)$ th ULDA discriminant vector φ_{i+1} , which maximizes the Fisher criterion function $J_F(\mathbf{W})$ with \mathbf{S}_t -orthogonal con-

straints, is the eigenvector corresponding to the maximum eigenvalue of the eigenequation: $\mathbf{U} \mathbf{S}_b \varphi = \lambda \mathbf{S}_w \varphi$, where

$$\mathbf{U} = \mathbf{I} - \mathbf{S}_t \mathbf{D}^T (\mathbf{D} \mathbf{S}_t \mathbf{S}_w^{-1} \mathbf{S}_t \mathbf{D}^T)^{-1} \mathbf{D} \mathbf{S}_t \mathbf{S}_w^{-1}, \quad \mathbf{D} = [\varphi_1 \varphi_2 \dots \varphi_i]$$

and \mathbf{I} is the identity matrix.

2.2. UHLDA based on the weighted pairwise Chernoff criterion

Although the set of discriminant vectors obtained by the ULDA technique can ensure the transformed feature components to be uncorrelated, which may significantly favor the subsequent pattern recognition tasks, these discriminant vectors set are actually not optimal due to the deficiencies described in introduction. Here, we introduce the multi-class Chernoff criterion function $J_C(\mathbf{W})$ into the ULDA, which can be regarded as the heteroscedastic extension of the Fisher criterion $J_F(\mathbf{W})$. The Chernoff criterion-based LDA solution was derived by Loog [1] and its efficiency has been experimentally demonstrated. However, the incorporation of the Chernoff criterion within the ULDA framework can ensure the discriminatory information in both class means' differences and class covariance matrices' differences to be extracted while the transformed feature components are statistically uncorrelated. The multi-class Chernoff criterion is defined as: $J_C(\mathbf{W}) = |\mathbf{W}^T \mathbf{S}_c \mathbf{W}| / |\mathbf{W}^T \mathbf{S}_w \mathbf{W}|$, where positive semi-definite \mathbf{S}_c is the multi-class directed distance matrix that captures the summation of Chernoff distances between different class pairs and is defined by

$$\begin{aligned} \mathbf{S}_c &= \sum_{i=1}^C \sum_{j=i+1}^C P_i P_j \mathbf{S}_C^{ij} \\ &= \sum_{i=1}^{C-1} \sum_{j=i+1}^C P_i P_j \mathbf{S}_w^{1/2} ((\mathbf{S}_w^{-1/2} \mathbf{S}_{ij} \mathbf{S}_w^{-1/2})^{-1/2}) \\ &\quad \times \mathbf{S}_w^{-1/2} (\mathbf{m}_i - \mathbf{m}_j) (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_w^{-1/2} (\mathbf{S}_w^{-1/2}) \\ &\quad \times \mathbf{S}_{ij} \mathbf{S}_w^{-1/2})^{-1/2} \\ &\quad + \frac{1}{\pi_i \pi_j} (\log(\mathbf{S}_w^{-1/2} \mathbf{S}_{ij} \mathbf{S}_w^{-1/2}) - \pi_i \log(\mathbf{S}_w^{-1/2} \mathbf{S}_i \mathbf{S}_w^{-1/2}) \\ &\quad - \pi_j \log(\mathbf{S}_w^{-1/2} \mathbf{S}_j \mathbf{S}_w^{-1/2})) \mathbf{S}_w^{1/2}, \end{aligned} \quad (2)$$

where $\pi_i = P_i / (P_i + P_j)$ and $\pi_j = P_j / (P_i + P_j)$ are relative a priori taking into account two classes that define the particular pairwise term. \mathbf{S}_C^{ij} and \mathbf{S}_{ij} are the pairwise directed distance matrix between classes i and j and the average pairwise within-class scatter matrix defined as $\pi_i \mathbf{S}_i + \pi_j \mathbf{S}_j$. \mathbf{S}_i and \mathbf{S}_j are the covariance matrices of classes i and j , respectively. The detailed derivation is presented in Ref. [1].

From Eq. (2), we notice that, in \mathbf{S}_c , all class pairs have same weights irrespective of their separability in the original space, which may possibly yield bad discriminant directions favoring originally well-separated class pairs. In order to avoid this problem, we introduce a weighting factor $w_{ij}(d_{ij}^c)$

for each class pair. As suggested by Loog [1], this weighting factor is chosen as

$$w_{ij}(d) = \frac{1}{2d^2} \operatorname{erf} \left(\frac{d}{2\sqrt{2}} \right) \quad \text{where } \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

is related to the pairwise approximated Bayesian accuracy. Therefore, the neighboring class pairs, which are not well separated in the original feature space, can have increased influence in the computation of \mathbf{S}_C than the class pairs well separated in the original feature space. We can modify the multi-class directed distance matrix \mathbf{S}_C , defined in Eq. (2), by introducing weighting factors as follows:

$$\tilde{\mathbf{S}}_C = \sum_{i=1}^C \sum_{j=i+1}^C P_i P_j w_{ij}(d_{ij}^c) \mathbf{S}_{ij}^c, \quad (3)$$

where

$$d_{ij}^c = \frac{\pi_i \pi_j}{2} (\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_{ij}^{-1} (\mathbf{m}_i - \mathbf{m}_j) + \frac{1}{2} (\log |\mathbf{S}_{ij}| - \pi_i \log |\mathbf{S}_i| - \pi_j \log |\mathbf{S}_j|)$$

is the pairwise Chernoff distance measure.

By substituting the between class scatter matrix \mathbf{S}_b used in the ULDA (Section 2.1) with the multi-class pairwise weighted directed distance matrix $\tilde{\mathbf{S}}_C$ (Eq. (3)) and following the same steps as the ULDA to generate discriminant vectors, we derive a novel UHLDA technique. The UHLDA increases the upper bound on the number of extracted discriminant vectors from $C - 1$ (ULDA) to $N - 1$, where N is the dimension of the original feature space, and all discriminatory information existing in the differences between all pairwise class means and pairwise class covariance matrices can be successfully extracted by properly adapting the weights of well and poorly separated class pairs in the original feature space. Furthermore, all the transformed feature components are statistically uncorrelated as in the ULDA technique.

It is worth noting that although the ULDA solution has been proven to coincide with the LDA solution in Ref. [3], we find that this conclusion is true only if the equation $\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_b$ is satisfied, which is based on the traditional definition of \mathbf{S}_b , \mathbf{S}_w and \mathbf{S}_t in the Fisher criterion. Here \mathbf{S}_t is the total scatter matrix estimated as $\mathbf{S}_t = \sum_{i=1}^N (\mathbf{X}_i - \mathbf{m})(\mathbf{X}_i - \mathbf{m})^T$. In our UHLDA, due to integrating the new weighted pairwise Chernoff criterion to replace the original Fisher criterion, i.e. we change \mathbf{S}_b into another matrix $\tilde{\mathbf{S}}_C$, \mathbf{S}_t is not equal to $\mathbf{S}_w + \tilde{\mathbf{S}}_C$. Therefore, the traditional LDA solution based on the proposed weighted pairwise Chernoff criterion cannot coincide with the ULDA solution based on the same criterion, and only the proposed UHLDA technique that combines the ULDA with the weighted pairwise Chernoff criterion can obtain the statistically uncorrelated feature components as verified by our experiments.

2.3. Enhancing numerical stability of the UHLDA

Although the UHLDA has demonstrated promising characteristics by extracting more powerful discriminant vectors, it may possibly suffer from some instability problems during the implementation. Here, we present the following solutions to enhance the UHLDA technique:

1. Due to the usage of the classwise covariance matrix \mathbf{S}_i , $1 \leq i \leq C - 1$, in the calculation of $\tilde{\mathbf{S}}_C$ (Eq. (3)), the singularity of \mathbf{S}_i may possibly occur when the number of training samples in some classes is too small even if the average class within class scatter matrix \mathbf{S}_w is nonsingular (e.g. E-coli data in our experiment). To overcome this problem, we employ a recently proposed covariance estimation technique called maximum entropy covariance selection method [5], which forms a new covariance matrix \mathbf{S}_i^{new} by considering the combination of both original covariance matrix \mathbf{S}_i and the average within-class scatter matrix \mathbf{S}_w (assumed to be nonsingular) based on the maximum entropy principle. The calculation of the estimated covariance matrix \mathbf{S}_i^{new} is shown in Table 1.

During the calculation of $\tilde{\mathbf{S}}_C$, if \mathbf{S}_i becomes singular, we can use the new covariance estimator \mathbf{S}_i^{new} to substitute the original one and thus the singularity problem of individual class covariance matrix can be solved.

2. In our implementation, we use the singular value decomposition (SVD) technique, instead of the eigendecomposition, to compute the eigenvectors, eigenvalues, inversion and logarithm.

3. Experimental results

We test the performance of our UHLDA technique on 4 data sets from the UCI repository: Ionosphere, Sonar, Ecoli¹ and Pendigits. Compared with the ULDA solution, the UHLDA demonstrates its superiority by extracting more discriminatory features, thereby improving the final classification results.

We employ the Bayesian linear discriminant classifier (LDC) and quadratic discriminant classifier (QDC) to evaluate the discriminating power of the extracted features. For the first 3 data sets, classification rate is based on the 10-fold cross-validation. The classification rate for the last data set is based on the given separate test set. In our experiments, we first remove the null space $\mathbf{S}_t(\mathbf{0}) = \{\mathbf{X} | \mathbf{S}_t \mathbf{X} = \mathbf{0}\}$ of the total scatter matrix \mathbf{S}_t due to the fact that the null space of $\mathbf{S}_t(\mathbf{0})$ does not include any discriminatory information. Then, the UHLDA is performed in the orthogonal complement space of $\mathbf{S}_t(\mathbf{0})$.

Table 2 shows the best classification results and the corresponding optimal dimensionality (shown in the

¹The E-coli data set includes some classes with training samples smaller in number than the data dimension, where the singularity problem occurs.

Table 1
The maximum entropy covariance selection algorithm

1. Find the eigenvector set $\Phi = [\phi_1, \phi_2, \dots, \phi_N]$ of the combined covariance given by $S_i + S_w$.
2. Calculate the variance contribution of both S_i and S_w under the base of Φ , i.e. calculate $[\zeta_1^i, \zeta_2^i, \dots, \zeta_N^i] = \text{diag}(\Phi^T S_i \Phi)$ and $[\zeta_1^w, \zeta_2^w, \dots, \zeta_N^w] = \text{diag}(\Phi^T S_w \Phi)$.
3. Form a new covariance estimator $S_i^{new} = \Phi \text{diag}[\max(\zeta_1^i, \zeta_1^w), \dots, \max(\zeta_N^i, \zeta_N^w)] \Phi^T$.

Table 2
Classification performance based on the ULDA and UHLDA techniques

| Data sets | Methods | Ionosphere (315/34/2) | Sonar (208/60/2) | Ecoli (336/7/8) | Pendigits (7494/[3498]/16/10) |
|-----------|----------|---------------------------------------|---------------------------------------|--------------------------------------|-------------------------------|
| LDC | ULDA | 0.8689 (1) ±0.0650 | 0.7362 (1) ±0.0955 | 0.8750 (6) ±0.0519 | 0.8296 (6) |
| | UHLDA | 0.8775 (10) ± 0.0633 | 0.7600 (8) ± 0.1078 | 0.8781 (6) ± 0.0561 | 0.8330 (10) |
| | Original | 0.8689 ±0.0650 | 0.7362 ±0.0955 | 0.8690 ±0.0562 | 0.8296 |
| QDC | ULDA | 0.8602 (1) ±0.0565 | 0.7362 (1) ±0.0955 | 0.8393 (6) ±0.0246 | 0.9128 (9) |
| | UHLDA | 0.9087 (6) ± 0.0536 | 0.7750 (14) ± 0.0856 | 0.8661 (5) ± 0.0346 | 0.9605 (15) |
| | Original | 0.8773 ±0.0574 | 0.7607 ±0.0828 | 0.8038 ±0.0660 | 0.9591 |

Note: The numbers in parentheses below the names of data sets represent (No. of training samples [No. of testing samples]/data dimensionality/total number of pattern classes).

parentheses) obtained by the ULDA and UHLDA techniques based on the LDC and QDC classification rates, respectively. The original classification results without dimensionality reduction are also presented. We can observe that classification rates based on the extracted features by our UHLDA are the highest. Especially for the QDC, the UHLDA solution achieves significant improvements over the ULDA solution since both the UHLDA and QDC methods take into account the second-order information, i.e. the differences between per class covariance matrices. It is also clear that the UHLDA technique can extract more discriminant vectors without the limitation of $C - 1$.

References

- [1] M. Loog, R.P.W. Duin, Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion, *IEEE Trans. Pattern Analysis and Machine Intelligence* 26 (6) (2004) 732–739.

- [2] Z. Jin, J.Y. Yang, Z.S. Hu, Z. Lou, Face recognition based on the uncorrelated discriminant transformation, *Pattern Recognition* 34 (7) (2001) 1405–1416.
- [3] Z. Jin, J.Y. Yang, Z.M. Tang, Z.S. Hu, A theorem on the uncorrelated optimal discriminant vectors, *Pattern Recognition* 34 (10) (2001) 2041–2047.
- [4] J. Yang, J.Y. Yang, D. Zhang, What's wrong with Fisher criterion?, *Pattern Recognition* 35 (11) (2002) 2665–2668.
- [5] C.E. Thomaz, D.F. Gillies, R.Q. Feitosa, A new covariance estimate for Bayesian classifiers in biometric recognition, *IEEE Trans. Circuits and Systems for Video Technology, (special Issue on Image- and Video-Based Biometrics)* 14 (2) (2004) 214–223.