Efficient Feature Extraction Based on Regularized Uncorrelated Chernoff Discriminant Analysis

A. K. Qin¹, P. N. Suganthan¹ and M. Loog²

¹School of EEE, Nanyang Technological University, Singapore, 639798
²The Image Group, IT University of Copenhagen, Rued Langgaards Vej 7, 2300, Copenhagen S, Denmark
qinkai@pmail.ntu.edu.sg, epnsugan@ntu.edu.sg, marco@itu.dk

Abstract

In this paper, two Regularized Uncorrelated Chernoff Discriminant Analysis (RUCDA) techniques are introduced. As a heteroscedastic extension of the class-wise weighted Fisher criterion, the class-wise weighted Chernoff criterion employed in RUCDA better approximates the Chernoff upper bound of the Bayes classification error in the transformed space, which enable the resulting RUCDA to extract uncorrelated discriminatory information from both mean and covariance differences. Experiments performed on UCI benchmark and protein secondary structure datasets demonstrate good performance of the proposed technique.

1. Introduction

Fisher’s linear discriminant analysis (FLDA) [1] seeks an optimal set of discriminant vectors \( \mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_c] \) to map the original \( n \)-dimensional feature space \( \mathbb{R}^n \) onto a lower dimensional feature space \( \mathbb{R}^m (m < n) \), by maximizing the Fisher criterion: \( J_F(\mathbf{W}) = tr(\mathbf{W}^T \mathbf{S}_b \mathbf{W}) / tr(\mathbf{W}^T \mathbf{S}_w \mathbf{W}) \). Here, \( \mathbf{S}_b \) and \( \mathbf{S}_w \) are the between-class and average within-class scatter matrices of the training sample vectors respectively, and are estimated as follows:

\[
\mathbf{S}_b = \sum_{i=1}^{C} P_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T = \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} P_i P_j (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T
\]

\[
\mathbf{S}_w = \sum_{i=1}^{C} P_i \mathbf{S}_i
\]

where \( C \), \( P_i \), \( \mathbf{m}_i \), \( \mathbf{m} \) and \( \mathbf{S}_i \) represent the number of pattern classes, a priori probability of pattern class \( \omega_i \), the mean vector of samples in class \( \omega_i \), the mean vector of all samples and the covariance matrix of samples in class \( \omega_i \), respectively. The total scatter matrix equals \( \mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w \).

Although FLDA’s success has been undoubtedly demonstrated in various fields, its intrinsic homoscedastic property may degrade its performance when dealing with data having different covariance matrices for different classes. Secondly, as indicated in [2], maximizing the Fisher criterion employed in FLDA does not directly correspond to the minimization of the Bayes classification error in the transformed space. Thirdly, some solutions to maximizing the Fisher criterion, e.g. the Foley-Sammon discriminant analysis [3], will lead to correlated feature components, which may deteriorate the performance of the subsequent learning model. Therefore, uncorrelatedness of the extracted features would be beneficial. In the past decades, many improved variants of FLDA have been developed to tackle those problems. Heteroscedastic discriminant analysis approaches [4] such as the LDR approach based on Chernoff criterion, divergence criterion and so on, can extract the discriminatory information from the differences of covariance matrices among class pairs instead of only from that of class means as FLDA does. As a result, the number of extracted discriminant features can break the limit of \( c-1 \). The multi-class LDR approach based on the weighted pairwise Fisher criterion, proposed by Loog [2], adds a weighting term related to pairwise Bayes accuracy of each class pair in the pairwise decomposition formulation of \( \mathbf{S}_b \). The obtained weighted pairwise between-class scatter approximates the union of pairwise Bayes accuracy in the transformed space and thus better relates to the minimization of Bayes classification error. Many uncorrelated linear discriminant analysis (ULDA) methods [5]-[7] have been proposed to extract statically uncorrelated features.

In this paper, we present two Regularized Uncorrelated Chernoff Discriminant Analysis (RUCDA) techniques to effectively tackle the above problems. The Chernoff criterion used in RUCDA can overcome the homoscedastic limitation of the Fisher criterion. A recently developed regularization scheme called Maximum Entropy Covariance Selector (MECS) [8] method is incorporated in the RUCDA to avoid numerical insatiably. To make the maximization of the Chernoff criterion more closely related to the minimization of Bayes classification error, a weighting term related to the Chernoff upper error bound in terms of Bayes decision, named class-pair weight, is added to weight each pairwise
directed distance matrix in the formulation of multi-class directed distance matrix. Moreover, we introduce another within-class weight to suppress the adverse influence from possible outlier classes in estimating the average within-class scatter matrix. In addition, the RUCDA method inherits the uncorrelated property of ULDA to extract statistically uncorrelated features.

2. Class-wise weighted uncorrelated FDA

Uncorrelated features are usually desirable in pattern recognition tasks because the performance of some popular classifiers deteriorates when correlated features are provided (see [1]). Jin et al. [5] proposed the uncorrelated LDA technique (ULDA), which can obtain discriminant vectors by maximizing the Fisher criterion under the constraint that the extracted feature components are statistically uncorrelated. Ye et al [7] also devised several LDA techniques to obtain uncorrelated discriminatory vectors by virtue of the property \(s_b = s_w + s_a\) and simultaneous diagonalization of \(s_b\) and \(s_a\). Yang et al. [6] 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 [5],[6]. However, Fisher criterion employed in ULDA is not directly related to minimizing the Bayes classification error. It suffers from two major deficiencies: 1) it attempts to find a linear transformation to maximize the sum of squared class pair center distances in the projected space. Obviously, class pairs with large distances would dominate the eigen-decomposition of \(s_b\), such that the obtained discriminative directions attempt to preserve the distances of already well-separated classes while causing large overlap between pairs of classes that are not well separated in the original space. 2) The estimation of the average within-class scatter matrix \(s_w\) in Eq. (1) assumes that all classes have the same covariance matrix (homoscedasticity), which is usually violated in reality. Therefore, classes with largely deviating covariance matrices may dominate the eigen-decomposition of \(s_w\). If those dominated covariance matrices coincidently belongs to outlier classes far away from all other classes, the resulting projection will attempt to minimize the spread of the outlier classes while neglecting the minimization of the covariance matrices that indeed impacts the classification error.

To tackle the above deficiencies in ULDA, we devise a class-wise weighted ULDA (cwULDA) solution by introducing class-pair weights and within-class weights in the estimation of the between-class and within-class scatter matrices \(s_b\) and \(s_w\), respectively. For each class pair in the pairwise decomposed definition of \(s_b\) in Eq. (1), we introduce a class-pair weighting factor, \(w^c(i, j) = (\sqrt{2}d^2)erf(d/\sqrt{2})\), where \(erf(x) = (2/\sqrt{\pi})\int_0^x e^{-t^2} dt\), as suggested in [2] to control the contribution of individual class pairs to the overall estimation so as to make this estimation better approximates Bayesian accuracy in the transformed space. Thus \(\tilde{s}_b\) in Eq. (1) is modified to:

\[
\tilde{s}_b = \sum_{i=1}^C \sum_{j=1}^i \sum_{m=1} p_i p_j w^c(i, j) d_{ij} [(m_i - m_j)(m_i - m_j)^T]
\]

(2)

where \(d_{ij} = (m_i - m_j)^T s_w^{-1} (m_i - m_j)\) is the Mahalanobis distance between classes \(\omega_i\) and \(\omega_j\) in the original feature space. Since the within-class weighting term for each class is used to eliminate the influences from the possible outlier classes on estimating average within-class scatter matrix, it can be directly calculated as the summation of the class-pair weighting factors as: \(w^c(i, j) = \sum_{m=1}^C p_i p_j w^c(i, j)\) \(i = 1,...,C\).

Therefore, the outlier classes with small class-pair weighting factors to the other classes also receive small weights in the estimation of the weighted average within-class scatter matrix, i.e.

\[
\tilde{s}_w = \sum_{i=1}^C p_i w^c(i, i) \tilde{s}_i
\]

(3)

In this paper, two cwULDA versions are implemented based on both the sequential [5] and batch [1],[7] extraction schemes, which are named cwULDA1 and cwULDA2. In cwULDA1, the discriminant vectors \(\phi_i, i = 1,...,m\) are extracted one after another (details in [5]) by maximizing the modified class-wise weighted Fisher criterion \(J'(W)\) subjecting to the \(s_i\) -orthogonal constraints: \(\phi_i s_i \phi_j = 0, \forall i \neq j, i, j = 1,...,m\), where \(\phi_j\) are already extracted discriminant vectors. The maximal number of discriminant vectors corresponding to non-zero Fisher criterion value is \(C-1\) due to the fact \(\text{rank}(\tilde{s}_b) \leq C - 1\).

In cwULDA2, the total scatter matrix \(s_t\) is first whitened, i.e. \(A_t^{-1/2} P_t s_t P_t A_t^{-1/2} = I\). We then transform the feature space into the \(s_t\) whitening space, within which the weighted pairwise between-class scatter matrix \(s_b\) and weighted average within-class scatter matrix \(\tilde{s}_w\) are calculated by Eqs. (2) and (3). Here, instead of the commonly used Fisher criterion formulated as the Raleigh quotient expression, we employ a subtractive variant of it, given by \(J(W) = p_t W^T s_t W - W^T s_t s_t W\), which was introduced in [1] and recently rediscovered in [9] with the concept of Maximum Margin Criterion (MMC). The validity and effectiveness of \(J(W)\) has been demonstrated through experiments. The reasons for this choice are: 1) since the equality \(s_t = \tilde{s}_b + \tilde{s}_w\) no longer holds for the class-wise weighted matrices \(\tilde{s}_b\) and \(\tilde{s}_w\), its simultaneous diagonalization in the \(s_t\) whitening space cannot be simply achieved as Ye’s methods [7] do. Therefore, we
need to find an orthonormal transformation matrix \( P_2 \), which maximizes \( \tilde{S}_b \) and minimizes \( \tilde{S}_w \) while keeping \( \tilde{S}_i \) whitened. However, the solution of conventional Fisher criterion leads to eigenvalue decomposition problem: \( \tilde{S}_bP = \tilde{P}A \), and the obtained transformation matrix \( P \) usually does not conform to orthonormality. In order to find a suitable \( P_2 \), we hope to seek a symmetric matrix that can simultaneously lead to the maximization of between-class scatter and minimization of within-class scatter.

### 3. Class-wise weighted uncorrelated Chernoff discriminant analysis

To solve the homoscedastic restriction of the Fisher criterion, we propose a class-wise weighted Chernoff criterion \( J_c(W) \) to substitute the class-wise weighted Fisher criterion \( J_f(W) \) used in cwULDA. The Chernoff criterion \( J_c(W) \) based LDA solution was derived by Loog [4]. The incorporation of the Chernoff criterion into the UDLA framework can ensure the discriminant information in both class mean and covariance’s differences to be extracted while the obtained feature components are statistically uncorrelated. The multi-class Chernoff criterion is: \( J_c(W) = tr(W^T S_b W)/tr(W^T S_w W) \), where positive semi-definite matrix \( S_b \) is the multi-class directed distance matrix that captures the sum of Chernoff distances between different class pairs, i.e.

\[
S_b = \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} PP_j S_{ij} = \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} PP_j \left( S_{ij} - \frac{1}{2}(m_i - m_j)(m_i - m_j)^T \right)
\]

\[
(S_{ij}) = \frac{1}{2} \left( \frac{1}{\pi \sigma_i \sigma_j} \log(S_{ij}) - \frac{1}{2} \log(S_i) - \frac{1}{2} \log(S_j) \right)
\]

where \( \pi_i = P_i/(P_i + P_j) \) and \( \pi_j = P_j/(P_i + P_j) \) are relative \textit{a priori} taking into account two classes that define the particular pairwise term \( S_{ij} \), and \( S_{ij} \) are the pairwise directed distance matrix between classes \( o_i \) and \( o_j \). The average pairwise within-class scatter matrix is defined as \( \pi_i S_i + \pi_j S_j \). The \( S_i \) and \( S_j \) are covariance matrices of classes \( o_i \) and \( o_j \), respectively.

The discriminative transformation obtained by maximizing the above Chernoff criterion is also a suboptimal solution to the maximization of Bayesian classification accuracy in the transformed space due to the same reason as described in section 2. Therefore, we also introduce the class-pair weights and within-class weights into \( S_b \) and \( S_w \). The resulting class-weighted Chernoff criterion would be more closely related to the optimal Bayes classification. Under the assumption that the covariance matrices for all the classes are equal in the Fisher criterion, the Bayes classification error can be explicitly expressed, which results in a suitable class-pair weighting factor to enable the class-wise weighted Fisher criterion to better approximate the Bayes classification accuracy. However, for heteroscedastic data, exact Bayes classification error between two classes cannot be derived while the upper bound of this error can be calculated as the function of the Chernoff distance \( D_0 \), i.e. the upper bound of error can be calculated as

\[
UpErr = \left[ \frac{1}{D_0^2} \left( \frac{P_i}{P_i + P_j} \right) \right] \exp \left( - \frac{D_0^2}{2} \right)
\]

and the resulting class-weighted Chernoff discriminant analysis counterparts, which can increase the upper bound on the number of extracted discriminant vectors from \( C-1 \) as in cwULDA to \( n-1 \) and all uncorrelated discriminatory information present in differences between class means and covariance matrices can be extracted.

### 4. Regularized Chernoff discriminant analysis

Due to using the class-wise covariance matrix \( S_{i}, 1 \leq i \leq C-1 \) in the calculation of \( \tilde{S}_C \) (Eq. 5), the singularity of \( S_i \) may possibly occur when the number of training samples in some classes is too small even if the average class within-class scatter matrix \( S_w \) is nonsingular. To overcome this problem, we employ a recently proposed covariance estimation technique MECS [8]. We call the resulting technique as Regularized Chernoff Discriminant Analysis (RCDA).

### 5. Experimental results

We test the performance of the cwULDA1, cwULDA2, RCDA1 and RCDA2 methods on 6 UCI datasets and 1 protein secondary structure prediction dataset [10]: (a). Bupa (345/6/2), (b). Ionosphere (351/34/2), (c). Sonar

---

0-7695-2521-0/06/$20.00 (c) 2006 IEEE
(208/60/2), (d), Iris (150/4/3), (e). Wine (178/13/3), (f). Vehicle (846/18/4) and (g). Protein (698/125/4) (#samples/#dimensions/#classes). Compared with ULDA solutions, the proposed approaches demonstrate superiority by extracting more discriminatory features, thereby improving the final classification results.

We employed linear and quadratic discriminant classifiers to evaluate discriminating power of the extracted features. For all datasets normalized to 0 mean and unit variance, classification rate is averaged over 10-fold cross-validation. The null space of $S_0$ is hereby removed since it contains no discriminatory information.

Tables 1 and 2 show the best classification results and the corresponding optimal dimensionality (shown in the parentheses) obtained by $M$: ULDA1, $M_2$: ULDA2, $M_3$: cwULDA1, $M_4$: cwULDA2, $M_5$: RCDA1 and $M_6$: RCDA2 methods based on the LDC and QDC classifiers, respectively. The original classification results without dimensionality reduction are also presented. We can observe that classification rates based on the extracted features by the RCDA techniques are the highest. Especially for the QDC, the RCDA solutions achieve much more improvements over other solutions since both the RCDA and QDC methods take into account the second order information. For ULDA techniques, the weighted versions cwULDA will coincide with the non-weighted versions ULDA. The RCDA techniques can extract discriminant vectors more than C-1. Note that the optimal accuracy and corresponding indiscernible values based on the Wilcoxon signed rank test with significant level 0.01 are typeset in bold.

**Table 1. Average accuracy by LDC classifier**

<table>
<thead>
<tr>
<th>D</th>
<th>Org</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
<th>M6</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>71.03</td>
<td>69.87</td>
<td>69.87</td>
<td>69.87</td>
<td>69.87</td>
<td>71.33</td>
<td>71.03</td>
</tr>
<tr>
<td></td>
<td>4.85</td>
<td>6.12</td>
<td>6.12</td>
<td>6.12</td>
<td>6.12</td>
<td>5.33</td>
<td>5.33</td>
</tr>
<tr>
<td>b</td>
<td>85.18</td>
<td>84.67</td>
<td>84.67</td>
<td>84.67</td>
<td>84.67</td>
<td>85.47</td>
<td>85.47</td>
</tr>
<tr>
<td></td>
<td>2.64</td>
<td>2.73</td>
<td>2.73</td>
<td>2.73</td>
<td>2.73</td>
<td>3.68</td>
<td>3.43</td>
</tr>
<tr>
<td>c</td>
<td>73.98</td>
<td>73.50</td>
<td>73.50</td>
<td>73.50</td>
<td>73.50</td>
<td>74.45</td>
<td>74.93</td>
</tr>
<tr>
<td></td>
<td>5.51</td>
<td>5.50</td>
<td>5.50</td>
<td>5.50</td>
<td>5.50</td>
<td>5.91</td>
<td>6.25</td>
</tr>
<tr>
<td>d</td>
<td>98.88</td>
<td>98.33</td>
<td>99.44</td>
<td>98.33</td>
<td>99.44</td>
<td>99.44</td>
<td>98.89</td>
</tr>
<tr>
<td></td>
<td>2.34</td>
<td>2.68</td>
<td>1.76</td>
<td>2.68</td>
<td>1.76</td>
<td>1.76</td>
<td>2.34</td>
</tr>
<tr>
<td>e</td>
<td>98.00</td>
<td>98.00</td>
<td>98.00</td>
<td>98.00</td>
<td>98.00</td>
<td>98.00</td>
<td>98.00</td>
</tr>
<tr>
<td></td>
<td>3.22</td>
<td>3.22</td>
<td>3.22</td>
<td>3.22</td>
<td>3.22</td>
<td>3.22</td>
<td>3.22</td>
</tr>
<tr>
<td>f</td>
<td>78.13</td>
<td>78.02</td>
<td>78.73</td>
<td>78.01</td>
<td>74.24</td>
<td>78.49</td>
<td>78.49</td>
</tr>
<tr>
<td></td>
<td>2.41</td>
<td>2.35</td>
<td>3.17</td>
<td>2.35</td>
<td>4.47</td>
<td>3.33</td>
<td>3.24</td>
</tr>
<tr>
<td>g</td>
<td>81.24</td>
<td>81.24</td>
<td>81.24</td>
<td>81.24</td>
<td>81.52</td>
<td>82.10</td>
<td>82.38</td>
</tr>
<tr>
<td></td>
<td>4.23</td>
<td>3.41</td>
<td>2.34</td>
<td>3.24</td>
<td>2.57</td>
<td>2.53</td>
<td>3.83</td>
</tr>
</tbody>
</table>

**Table 2. Average accuracy by QDC classifier**

<table>
<thead>
<tr>
<th>D</th>
<th>Org</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
<th>M6</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>62.36</td>
<td>64.08</td>
<td>64.08</td>
<td>64.08</td>
<td>64.08</td>
<td>68.71</td>
<td>66.98</td>
</tr>
<tr>
<td></td>
<td>7.89</td>
<td>5.16</td>
<td>5.16</td>
<td>5.16</td>
<td>5.16</td>
<td>6.26</td>
<td>6.05</td>
</tr>
</tbody>
</table>

6. Future Works

The uncorrelated features extracted by the proposed RUCDA techniques may possibly facilitate the further feature weighting to enhance the subsequent learning models. Currently, some weighting schemes based on evolutionary computation techniques are being investigated. Moreover, extensive evaluation on high-dimensional bioinformatics datasets is also ongoing.

**Reference**