



PERGAMON

Available at  
www.ElsevierComputerScience.com

POWERED BY SCIENCE @ DIRECT®

Pattern Recognition 38 (2005) 485–493

PATTERN  
RECOGNITION

THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

# Linear dimensionality reduction using relevance weighted LDA

E.K. Tang<sup>a</sup>, P.N. Suganthan<sup>a,\*</sup>, X. Yao<sup>b</sup>, A.K. Qin<sup>a</sup>

<sup>a</sup>School of Electrical and Electronic Engineering Nanyang Technological University, Nanyang Avenue, Block S2, Singapore 639798, Singapore

<sup>b</sup>School of Computer Science, University of Birmingham, Birmingham, B15 2TT, United Kingdom

Received 23 June 2004; received in revised form 27 September 2004; accepted 27 September 2004

## Abstract

The linear discriminant analysis (LDA) is one of the most traditional linear dimensionality reduction methods. This paper incorporates the inter-class relationships as relevance weights into the estimation of the overall within-class scatter matrix in order to improve the performance of the basic LDA method and some of its improved variants. We demonstrate that in some specific situations the standard multi-class LDA almost totally fails to find a discriminative subspace if the proposed relevance weights are not incorporated. In order to estimate the relevance weights of individual within-class scatter matrices, we propose several methods of which one employs the evolution strategies.

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

**Keywords:** Feature extraction; Linear discriminant analysis; Weighted LDA; Evolution strategies; Approximate pairwise accuracy criterion; Chernoff criterion; Mahalanobis distance

## 1. Introduction

When solving a pattern classification problem, it is common to apply a feature extraction method as a pre-processing technique, not only to reduce the computation complexity, but possibly also to obtain better classification performance by reducing irrelevant and redundant information in the data. A class of feature extraction procedures can be defined by a transformation  $\mathbf{Y} = T(\mathbf{X})$ , where  $\mathbf{X} \in R^D$ ,  $\mathbf{Y} \in R^d$  and the transformation  $T$  is obtained by optimizing suitable objectives. Hence, the feature extraction can be considered to

have two parts, namely formulating suitable objectives and determining the corresponding optimal solution of  $T$ .

Linear discriminant analysis (LDA) is one of the most well-known linear dimensionality reduction (LDR) algorithms: Given a  $D$ -dimensional data set  $\mathbf{X}$  ( $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ ) consisting of  $C$  classes, the original Fisher criterion function applied in the LDA is,

$$J = \frac{\omega \mathbf{S}_B \omega^T}{\omega \mathbf{S}_W \omega^T} \quad (1)$$

where  $\mathbf{S}_W$  is the average within-class scatter matrix and  $\mathbf{S}_B$  is the between-class covariance matrix of  $\mathbf{X}$ . By performing an eigenvalue decomposition of  $\mathbf{S}_W^{-1} \mathbf{S}_B$  and taking the rows of  $\omega$  to equal the  $d$  eigenvectors corresponding to the  $d$  ( $d \leq C - 1$ ) largest eigenvalues, a  $d$ -by- $D$  ( $d \leq D$ ) transformation matrix  $\omega$  is determined such that the Fisher criterion of between-class scatter over average within-class scatter is maximized [1]. The data set  $\mathbf{X}$  can be mapped from the  $D$ -dimensional original space to the new  $d$ -dimensional

\* Corresponding author. Tel.: 0065 6790 5404; fax: 0065 6792 0415.

E-mail addresses: tangke@pmail.ntu.edu.sg (E.K. Tang), epnsugan@ntu.edu (P.N. Suganthan), X.Yao@cs.bham.ac.uk (X. Yao), qinkai@pmail.ntu.edu.sg (A.K. Qin)

URL: <http://www.ntu.edu.sg/home/EPNSugan>.

transformed space by the function  $Y = \omega X$ . It is believed that the new space preserves most of the discriminative information contained in  $\mathbf{X}$ , and hence good classification accuracy can still be achieved in this lower dimensional space. Therefore, LDA has been widely applied to solve problems such as face recognition [2], bioinformatics and medical image processing [3].

In spite of its popularity, LDA does not guarantee to find the optimal subspace in some situations. The solution of LDA is statistically optimal only when the distributions of samples in different classes satisfy the so-called homoscedastic Gaussian (HOG) model, i.e. the probability distribution functions (PDF) of samples in different classes obey Gaussian distributions, with distinct mean vectors but with the same covariance matrix for all the classes [4]. This assumption can seldom be satisfied in real-world problems, as the criterion involves the sample estimates of mean vectors and covariance matrices. The sample estimates are usually not identical to the true mean vectors and covariance matrices because of the lack of sufficient samples and/or the existence of outliers. Therefore, numerous modified versions of the LDA were proposed. To make the LDA work well for data sets that do not satisfy the HOG model, Kumar and Andreou [5] proposed a maximum likelihood approach called heteroscedastic discriminant analysis to remedy situations where the covariance matrices of the Gaussian distributed classes are not equal and not proportional. Recently, Loog et al. proposed another heteroscedastic extension to the LDA [6]. Proposed by Hastie et al., the mixture discriminant analysis fits each class using Gaussian mixture model, which works well for clustered data [7]. Moreover, various regularization techniques have been discussed to estimate the covariance matrices more accurately [8,9]. To tackle problems with small number of samples, such as the face recognition problem, combinations of PCA and LDA have also been continuously proposed [10,11]. In Liu’s and Wechsler’s work [11], several techniques were integrated so that the latent noisy information contained in the subspace corresponding to the relatively small eigenvalues of  $\mathbf{S}_W$  would not be preserved. Both Loog et al. [12] and Lotlikar and Kothori [13] proposed algorithms to restrain the negative influence of the so-called outlier classes on the estimation of between class covariance matrices.

In this paper, we propose a novel outlier-class-resistant scheme that can estimate the within-class covariance matrix  $\mathbf{S}_W$  more accurately for classification purposes. Then we present an LDR algorithm, which combines our proposed scheme with an existing improved version of LDA [12]. To further improve the overall performance, we also employ an evolutionary search algorithm, namely the evolution strategies (ES) [14]. We observe that several evolutionary algorithms have been applied in the pattern recognition field to perform search efficiently [15–21].

The remainder of this paper is organized as follows: in Section 2, we briefly introduce two relevant previous variants of the LDA. In Section 3, we present our relevance-

weighted overall within-class covariance matrix estimation scheme (RWW), weighted multi-class LDR (WLDR) and evolutionary WLDR (EWLDR). Experimental results on several synthetic and UCI data sets [22] are presented in Section 3.3. The conclusions and discussions are presented in Section 4.

## 2. LDA, aPAC and fractional LDA

In a data set, distance or similarity relationships between pairs of classes are important information for classification. The distance or similarity relationships, usually acquired through statistical approaches such as Euclidean distance, Mahalanobis distance, etc., reflect how well two classes are separated in the feature space. When first introduced by Fisher, the LDA was proposed for two-class problems [23]. Hence, relationships were not considered at first until Rao generalized the LDA to multiple classes [24]. Generally, the between-class scatter matrix is computed using the following equation:

$$\mathbf{S}_B = \sum_{i=1}^{C-1} \sum_{j=i+1}^C p_i p_j (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T, \quad (2)$$

where  $C$  is the number of classes,  $\mathbf{m}_i, \mathbf{m}_j$  and  $p_i, p_j$  are the mean vector and a priori probability of classes  $i$  and  $j$ , respectively.

However, in multi-class LDA, the relationships between pairs of classes are likely to be different from one pair to another. The classes that are closer to each other are potentially more confusing and they should be given more attention during the feature extraction stage. Based on this observation, Loog et al. named the class that is distant from other classes as “outlier class” [12]. Since Eq. (2) is not directly related to classification accuracy and focuses equally to every pair of classes, the outlier classes may negatively influence the estimation of the overall between-class covariance matrix  $\mathbf{S}_B$ . Therefore, Loog et al. [12] proposed an extended criterion named approximate pairwise accuracy criterion (aPAC) by replacing Eq. (2) with

$$\mathbf{S}_B = \sum_{i=1}^{C-1} \sum_{j=i+1}^C L_{ij} p_i p_j (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T, \quad (3)$$

where the weights  $L_{ij}$ s are usually estimated based on relationships between classes  $i$  and  $j$ . In aPAC, the non-optimality of Fisher criterion is restrained. To relate Fisher’s criterion to classification accuracy, aPAC employs a Mahalanobis distance-based relationship estimation function to compute the weights. This approach has been proven to be at least as powerful as the original LDA in many situations, while outperforming the original LDA clearly when one or more classes in a data set are far away from the other classes [12].

Another algorithm that focuses on the between-class covariance matrix estimation is the Fractional-Step linear

discriminant analysis (F-LDA) [13]. By moving incrementally from the  $D$ -dimensional space to the  $d$ -dimensional space through an  $r$ -step ( $r$  is a parameter set in advance) iterative procedure, F-LDA yields better classification accuracy. Actually, F-LDA can be viewed as an improved version of the aPAC. To transform a data set from a  $D$ -dimensional space into a  $d$ -dimensional space, one can just first find the optimal  $D-1$  subspace by the aPAC, update the relationships between classes and re-compute the weights, then use those weights to find the  $D-2$  dimensional subspace, after  $D-d$  iterations, a  $d$ -dimensional space can be obtained.

It should be noted that both aPAC and F-LDA preprocess the data with a whitening transformation [12,13]. This transformation rotates and scales the coordinate axes to transform the average within-class scatter matrix  $\mathbf{S}_W$  of the training samples to be a  $D$ -by- $D$  identity matrix  $\mathbf{I}$ , so that the  $\mathbf{S}_B$  contains all the discriminative information.

### 3. Weighted linear dimensionality reduction algorithm

#### 3.1. Motivation

In the original LDA and its improved variants, the covariance matrices of each class  $i$  are estimated by the equation

$$\mathbf{S}_{CW_i} = \sum_{j=1}^K (x_j - \mathbf{m}_i)(x_j - \mathbf{m}_i)^T, \quad (4)$$

where  $\mathbf{S}_{CW_i}$  is the covariance matrix of class  $i$ ,  $K$  is the number of samples in class  $i$  and  $\mathbf{x}_j$  is a sample vector of class  $i$ . After estimating  $\mathbf{S}_{CW_i}$  for all the classes separately ( $i = 1, 2, \dots, C$ ), the overall within-class covariance matrix  $\mathbf{S}_W$  is computed as follows:

$$\mathbf{S}_W = \sum_{i=1}^C p_i \mathbf{S}_{CW_i}. \quad (5)$$

Since distributions of all the classes are assumed to have the same covariance matrix, the  $\mathbf{S}_W$  is employed to represent the classes' distribution. It is definitely appropriate when all the  $\mathbf{S}_{CW_i}$ 's are identical to each other. However, in real data sets, the  $\mathbf{S}_{CW_i}$ 's are likely to differ between different classes. Hence, the estimation of  $\mathbf{S}_W$  obtained using Eq. (5) is unlikely to be the best estimation with respect to the classification performance.

The motivation for introducing relevance weights during the computation of the overall within-class covariance matrix arises from the following consideration. When  $\mathbf{S}_{CW_i}$ 's of different classes are quite different from each other,  $\mathbf{S}_W$  computed by using Eq. (5) may not be appropriate with respect to classification accuracy, although from a statistical point of view Eq. (5) is an unbiased estimation for the overall  $\mathbf{S}_W$ . In a classification problem, what we need is an  $\mathbf{S}_W$  that can yield a new space to achieve high classification accuracy. Furthermore, the worst case scenario is that if some

elements in one  $\mathbf{S}_{CW_i}$  are much larger than the corresponding elements in other  $\mathbf{S}_{CW_i}$ 's, it will have a dominant influence when computing  $\mathbf{S}_W$ . In this situation, Eq. (5) will yield an  $\mathbf{S}_W$  that can only represent the dominant  $\mathbf{S}_{CW_i}$  well. If the class with "dominant"  $\mathbf{S}_{CW_i}$  is simultaneously an outlier class in the feature space, Eq. (5) would fail in estimating  $\mathbf{S}_W$  for improved classification, as the LDA would primarily focus on minimizing within-class scatter of the outlier class without considering the others, while the outlier class can be easily classified in the original space and hence does not need much consideration during the transformation.

#### 3.2. Relevance weighted within-class covariance matrix (RWW)

From Subsection 3.1, we can see that in addition to assigning different considerations to classes when estimating the between-classes covariance matrix  $\mathbf{S}_B$ , a weighting scheme should also be employed when estimating  $\mathbf{S}_W$ . To reduce the influence of outlier classes, we modify Eq. (5) as below:

$$\mathbf{S}_W = \sum_{i=1}^C p_i r_i \mathbf{S}_{CW_i} \quad (6)$$

where  $r_i$ 's are the relevance-based weights. By integrating  $r_i$  in Eq. (5), we intend to ensure that if class  $i$  is an outlier class, it only influences the estimated  $\mathbf{S}_W$  slightly. This is reasonable since if one class is well separated from the other classes in the data set, then whether the within-class covariance matrix of this class in the new space is compact or not will not have much influence on classification.

To calculate a class's separability with other classes, we define a straightforward weighting function:

$$r_i = \sum_{j \neq i} \frac{1}{L_{ij}}. \quad (7)$$

Here  $L_{ij}$  is defined as the dissimilarity between classes  $i$  and  $j$  or how well classes  $i$  and  $j$  are separated in the original space. The  $r_i$ 's will be normalized so that the largest one of them is 1. Although several dissimilarity measures have been proposed in the past, it is impossible to choose one of them as the best measure independent of the data set. In this paper, we consider five measures: Euclidean distance (ED), Mahalanobis distance (MD) estimated Bayesian classification accuracy function (BA), the weighting function  $f$  proposed in aPAC [12] and the Chernoff distance employed in [6]. These dissimilarity measures can be calculated as follows:

Euclidean distance (ED):

$$L_{ij} = \sqrt{(\mathbf{m}_i - \mathbf{m}_j)^T (\mathbf{m}_i - \mathbf{m}_j)},$$

Mahalanobis distance (MD):

$$L_{ij} = \sqrt{(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_W^{-1} (\mathbf{m}_i - \mathbf{m}_j)},$$

Bayesian accuracy (BA):

$$L_{ij} = 0.5 + \frac{1}{\sqrt{\pi}} \int_0^{\text{MD}} e^{-t^2} dt$$

Weighting function of aPAC:

$$L_{ij} = \frac{1}{2\text{MD}^2} \int_0^{\text{MD}/2\sqrt{2}} e^{-t^2} dt$$

Chernoff criterion:

$$L_{ij} = (\mathbf{m}_i - \mathbf{m}_j)^T (\alpha \mathbf{S}_{C_{W_i}} + (1 - \alpha) \mathbf{S}_{C_{W_j}}) (\mathbf{m}_i - \mathbf{m}_j) + \frac{1}{\alpha(1 - \alpha)} \log \frac{|\alpha \mathbf{S}_{C_{W_i}} + (1 - \alpha) \mathbf{S}_{C_{W_j}}|}{|\mathbf{S}_{C_{W_i}}|^\alpha |\mathbf{S}_{C_{W_j}}|^{1-\alpha}}$$

In the Chernoff criterion,  $\alpha$  is a parameter that should be set by users between [0,1]. By using any one of the dissimilarity measures, we can obtain a  $C$ -dimensional vector  $\mathbf{V}$ , each component of which corresponds to one class. This method can also be viewed as changing the original global classification problem into a local problem, in which we focus mainly on discriminating those classes that are close to each other and thus hard to classify correctly. However, it is different from the idea of local LDAs. In a local LDA algorithm, one should first determine which local region samples should belong to by using a clustering-like method. After the classes are divided into sub-classes by a clustering method, the within-sub-class covariance matrices of different classes may still be different to each other, and relationships between pairs of sub-classes may also be different. Hence, RWW scheme can be beneficially integrated in this variant as well.

A straightforward example will show that Eq. (6) can restrain the worst case scenario that we described in Subsection 3.1, while not influencing the algorithm adversely in situations where the worst case scenario does not exist. Suppose there are three classes  $C_1, C_2$  and  $C_3$  in a data set, with within-class scatter matrices  $\mathbf{S}_{C_{W_1}}, \mathbf{S}_{C_{W_2}}$  and  $\mathbf{S}_{C_{W_3}}$ , respectively, and  $\mathbf{S}_{C_{W_1}} = \mathbf{S}_{C_{W_2}}$ . The dissimilarity measures between each pair of classes are  $L_{12}, L_{13}$  and  $L_{23}$ ,  $L_{13} = L_{23} \gg L_{12}$ . If elements of  $\mathbf{S}_{C_{W_3}}$  are much larger and thus  $\mathbf{S}_{C_{W_3}}$  is dominant among the three matrices, Eq. (5) will yield an  $\mathbf{S}_W \cong \mathbf{S}_{C_{W_3}}$ . Hence, the minimization of the overall within-class scatter  $\mathbf{S}_W$  would be equivalent to minimizing the within-class scatter  $\mathbf{S}_{C_{W_3}}$  of the outlier class. By employing Eq. (6) and Eq. (7) ( $r_1 = r_2 \gg r_3$ ), an  $\mathbf{S}_W$  that is very similar to  $\mathbf{S}_{C_{W_1}}$  and  $\mathbf{S}_{C_{W_2}}$  will be yielded. If  $\mathbf{S}_{C_{W_3}}$  is not dominant (for example,  $\mathbf{S}_{C_{W_3}} \cong \mathbf{S}_{C_{W_1}}$ ), the  $\mathbf{S}_W$ 's yielded by Eqs. (5) and (6) will be similar, and hence there is no negative influence due to the relevance weighting.

### 3.3. Relevance-weighted linear dimensionality reduction algorithm (WLDR)

As mentioned in Section 2, aPAC and F-LDA employ a whitening transformation as a preprocessing procedure

[12,13]. The whitening transformation is, in nature, a linear transformation  $\mathbf{X}_T = \mathbf{X}_T$ , where  $\mathbf{T}$  is computed as  $\mathbf{S}_W^{-1/2}$ . After the transformation,  $\mathbf{S}_W$  of  $\mathbf{X}_T$  becomes an identity matrix  $\mathbf{I}$ . Hence this transformation actually transfers all the discriminative information of  $\mathbf{S}_W$  into  $\mathbf{S}_B$ . Since it has been argued that Eq. (6) is likely to yield an appropriate  $\mathbf{S}_W$  with more discriminative information for classification, it would be promising to combine our weighted overall within-class covariance matrix computation scheme with the variants employing a whitening transformation to obtain better results than each of the variants as follows. First, the weight  $L$  for every pairs of classes is computed by one of the five measures listed above, then the weight  $r$  for each class is calculated by Eq. (7), and finally, the between- and within-class covariance matrices computed using Eqs. (3) and (6) are used in the Fisher criterion Eq. (1) to find the optimal transformation matrix.

### 3.4. Evolving the relevance weights (EWLDR)

Although WLDR as well as aPAC [12] and some other variants [6,13] choose weighting functions to make the Fisher criterion more representative of the separability between classes, these weighting functions may not be optimal for all the data sets. These approaches are based on some estimation of the relationships between the classes in the original space. However, in classification problems, the effectiveness of a weighting method would be best described by its ability to yield high classification accuracy, rather than by how well it can represent the between-class relationships. Therefore, the estimated dissimilarity measures should be further tuned to yield high classification accuracy. In this work, we employ the evolution strategies (ES) algorithm to tune the dissimilarity measures  $L_{ij}$  used in Eqs. (3) and (7).

As an evolution-based optimization algorithm, evolution strategies can be viewed as a population-based variant of generate-and-evaluate algorithms [25]. It employs search operators such as mutation to generate new solutions and a fitness evaluation measure to determine the survival of solutions to the next generation. By employing ES, we can obtain the dissimilarity measures  $L_{ij}$  corresponding to a higher classification accuracy on the validation data set. The major steps of EWLDR are presented in Table 1.

#### 3.4.1. Fitness evaluation

In most evolutionary computation-based feature extraction methods, classification accuracy is employed as the fitness measure [17–20]. In order to improve the generalization performance, we use the classification rate on the cross-validation data set as the fitness measure.

#### 3.4.2. Mutation

Mutation in EWLDR is carried out in two steps, namely weight mutation and normalization. In the first step, the

Table 1  
The EWLDR algorithm

1. Generate an initial population of  $M$  weighting vectors. Five of them are generated using the dissimilarity measures we presented in Subsection 3.2. The others are randomly generated by linearly combining the first five vectors. Compute the fitness values of these  $M$  vectors
2. Use all initial vectors as parents to create  $n_b$  offspring vectors by Gaussian mutation. In order to maintain population diversity and to prevent premature convergence, another  $n_r$  vectors are generated by randomly combining the first five vectors as in step 1
3. Calculate the fitness of  $n_b + n_r$  vectors, and prune the population to  $M$  fittest vectors.
4. Go to the next step if the maximum number of generations has been reached. Otherwise, go to step 2
5. Choose the fittest one in the population in the last generation as the optimal weighting vector.

parent dissimilarity measures are mutated by the following mutation operation:

$$L'_{ij} = L_{ij} + N(0, 1), \quad (8)$$

where  $L_{ij}$  denotes the dissimilarity between classes  $i$  and  $j$ , and  $N(0, 1)$  denotes a Gaussian random variable with zero mean and unit standard deviation. Further, each vector is normalized so that the largest element of it is 1.

Some advanced ES algorithms employ self-adaptive step size or different mutation operators in the mutation procedure [25–27]. In our work, we employ a simple version of the ES. The  $L_{ij}$  is only a relative importance measure. For example,  $V_1 = [0.1, 0.1, 0.2]$  is the same as  $V_2 = [0.5, 0.5, 1]$ . Hence, vectors can be normalized without adversely influencing the search for the global optimum and we only need the ES to further enhance the good initial estimations for weight vectors obtained using various dissimilarity measures.

#### 4. Experimental results

First, we make use of synthetic data sets to demonstrate the benefits of the RWW explained in Subsection 3.2. Further (in Subsection 4.3), we use six UCI data sets [22] for comparative evaluation of LDA, aPAC, WLDR and EWLDR algorithms.

To compare with the aPAC, we employ Loog's function in [12] to calculate the  $L_{ij}$ 's, and then use it to calculate the weights for RWW as described in Subsection 3.2. All the continuous features are scaled within  $[0,1]$ . The classification accuracy is computed using the linear discriminant classifier employed by Loog et al. [12]. The parameters used in EWLDR are set to be the same for all the data sets: the population size  $M$  (25), the number of generations (300), the number of offspring  $n_b$  (50) and  $n_r$  (25).

##### 4.1. Synthetic data set 1

This 9-D synthetic data set consists of three classes and 100 samples per class, thus the prior probabilities are the same for all the classes. Classes 1 and 2 have the same multivariate Gaussian distribution. Class 3 has a totally different within-class scatter matrix  $S_{CW3}$ , elements of which are

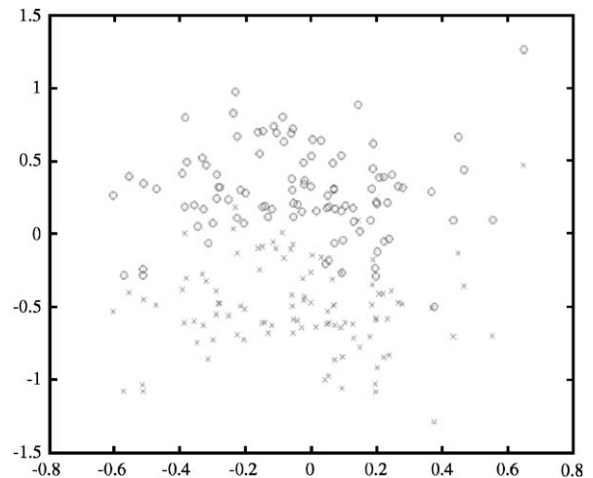


Fig. 1. Distribution of classes 1 and 2 in the new space with RWW.

much larger than that of  $S_{CW1}$  and  $S_{CW2}$ . To obtain such a situation, we randomly generate class 1, then move the whole of class 1 to generate class 2, so that  $S_{CW1} = S_{CW2}$ . The centers of class 1 and class 2 are near the origin and close to each other so as to have a little overlap between them, while the center of class 3 is at  $[10, 10]^T$ , so that it is an outlier class with the dominant  $S_{CW}$ .

Figs. 1 and 2 show the distribution of classes 1 and 2 after being transformed into a new 2-dimensional space with and without RWW scheme, respectively. Since class 3 is too far away from classes 1 and 2 and displaying it would make the figure size much larger, we did not show class 3 in the figures. From the two figures, we can observe that in this worst case situation, LDA makes the two classes that only overlap a little in the original space totally overlap each other in the transformed space, while by incorporating a weighting scheme they are kept separated.

##### 4.2. Synthetic data set 2

These synthetic data sets consists of eight 11-dimensional data sets, with 3–10 classes, respectively and 100 samples per class. In each data set, one class is set as outlier class, and the others overlap each other slightly. But within-class

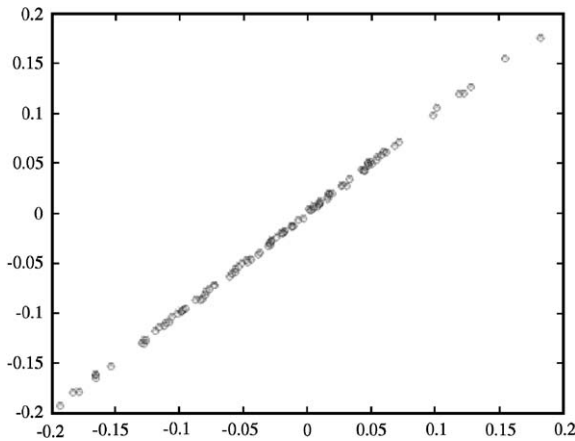


Fig. 2. Distribution of classes 1 and 2 in the new space without RWW.

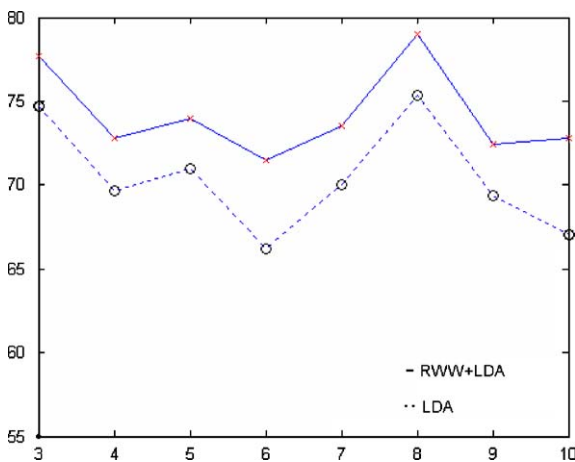


Fig. 3. Classification performance of RWW+LDA and LDA on synthetic data set 2. X-axis represents the number of classes in the data set, Y-axis represents the classification accuracy achieved on the extracted features.

scatter matrix of the outlier class is no longer dominant, so that the main goal of the weighting scheme here is to improve the LDA instead of preventing it from breaking down as in Section 4.1. For each data set,  $C-1$  features are extracted using the original LDA and the LDA with RWW scheme, respectively. The classification accuracies achieved on the extracted features are plotted in Fig. 3, which demonstrates that the RWW scheme can improve LDA when one or more outlier classes exist in the data set by estimating the  $S_W$  more appropriately for the classification task.

#### 4.3. Performance characterization using UCI data sets

To demonstrate the improvements that can be achieved by not only WLDR but also EWLDR over the variants em-

ploying a whitening transformation, we choose aPAC as an example, and use six UCI data sets to compare LDA, aPAC, WLDR, EWLDR. The six data sets are landsat, optdigits, vehicle, DNA, thyroid disease and vowel data sets.

*Landsat.* The Landsat data set is generated from landsat multi-spectral scanner image data. It has 36 dimensions, 4435 training samples and 2000 testing samples belonging to 6 classes.

*Optdigits.* This is a 60-dimensional data set on optical recognition of 10 handwritten digits. It has separate training and testing sets with 3823 and 1797 samples, respectively.

*Vehicle.* This data set involves classification of a given silhouette as one of four types of vehicles, consists of 846 samples, 18 features and 4 classes.

*DNA.* The DNA data set consists of separated training set and testing set, with 2000 and 1186 samples, respectively. It has 180 dimensions and 3 classes.

*Thyroid disease database.* This 21-dimensional data set has a separate training set, a separate testing set and 3 classes. The training set consists of 3772 samples and the testing set consists of 3428 samples.

*Vowel.* Vowel data set consists of 11 classes, and each class has 90 10-dimensional samples.

For data sets with separate training and testing samples, such as the Landsat, Optdigits, DNA and Thyroid, we train the algorithms with the training samples, and test them using the testing samples. For vehicle and vowel data sets, since separate testing sets are not available, 10-fold cross-validation is employed. Same as in Loog's work [12], for all six data sets, the original LDA, WLDR, aPAC and EWLDR are used to find subspaces from 1 to  $C-1$  dimensions, i.e., for landsat it is 1–5, for optdigits it is 1–9 and so on. Average and standard deviation of classification accuracies (CAs) on the testing sets of vehicle and vowel data sets are shown in Table 2. CAs on the other four data sets are shown in Table 3, since cross-validation is not employed for these data sets, no standard deviation is computed.

From the results, we can see that in all six data sets, both WLDR and EWLDR perform generally better than the aPAC and the LDA. As mentioned by Loog et al. [12], we observe that sometimes the original LDA performs the same or even better than aPAC and WLDR. The reason is that in these cases, a weighting vector computed by a deterministic function is not suitable with respect to classification. By employing evolution strategies, the weighting vectors are optimized to the particular classification problem. Therefore, the EWLDR performs the best in most of the cases. The improvement is especially obvious in lower dimensions. This demonstrates that our LDR algorithms can preserve more discriminative information in lower dimensional spaces.

## 5. Conclusions and discussion

This paper first introduces a novel outlier-class resistant scheme for estimating the overall within-class covariance

Table 2  
Classification results (%) of LDA, aPAC, WLDR and EWLDR on vehicle and vowel data sets

Data set	Dimensions	EWLDR	WLDR	aPAC	LDA
Vehicle	3	78.36/2.23	78.36/3.36	78.25/2.93	78.25/2.93
	2	75.88/1.38	75.53/1.14	75.06/2.52	75.06/2.29
	1	63.11/2.92	56.14/2.2	58.26/3.04	62.04/3.42
Vowel	10	62.12/3.23	62.12/3.23	62.12/3.23	62.12/3.23
	9	64.04/2.4	61.92/3.13	62.02/3.26	62.02/3.26
	8	64.14/2.94	61.72/3.65	61.72/3.63	61.92/3.5
	7	63.81/2.83	61.21/2.83	61.41/2.85	61.41/2.74
	6	63.54/4.18	62.53/3.82	62.42/3.77	62.73/3.6
	5	63.23/2.97	62.22/3.33	61.82/3.88	61.31/3.56
	4	62.93/2.26	60.2/2.9	60.4/3.46	60.61/3.67
	3	61.72/4.21	60.1/3.73	59.6/4.27	59.6/4.33
	2	61.62/3.91	60/3.73	59.29/4.21	59.6/4.47
	1	49.6/3.21	32.63/2.71	33.43/4.92	36.46/1.72

Table 3  
Classification results (%) of LDA, aPAC, WLDR, and EWLDR on validation set of landsat, optdigits, DNA and thyroid disease data sets

Data set	Dimensions	EWLDR	WLDR	aPAC	LDA
DNA	2	94.18	94.01	94.1	94.1
	1	82.04	81.03	80.27	76.98
Thyroid	2	93.99	93.93	93.87	93.87
	1	94.22	93.87	93.87	93.82
Landsat	5	83.15	82.9	82.65	82.65
	4	83.15	82.9	82.8	82.5
	3	82.35	81.95	82	81.85
	2	80.3	73.45	75.25	74.45
Optdigits	1	71.45	63.4	56.65	52.35
	9	93.88	93.93	93.88	93.88
	8	94.1	93.82	93.82	93.77
	7	91.93	92.93	92.82	92.65
	6	92.93	92.77	92.65	92.49
	5	91.26	90.76	90.6	90.04
	4	88.98	88.26	87.87	87.81
	3	83.23	82.25	82.53	79.8
	2	73.7	64.22	64.66	62.77
1	47.13	40.96	40.79	36.17	

matrix for the well-known linear discriminant analysis. Considering the fact that the minimization of within-class scatter of different classes has different importance in a classification problem, the novel scheme is obtained by introducing weights to within-class covariance matrices,  $\mathbf{S}_{CW}$ 's of each class when estimating the overall within-class scatter matrix  $\mathbf{S}_W$  that is used to obtain the transformation matrix  $\omega$ . The proposed method has two important properties. Firstly, it is stable in specific situations where the LDA would totally break down. Secondly, it can improve the performance of the LDA in general.

Although we only experimented on combining the RWW with the aPAC, RWW can also improve the performance of

F-LDA [13]. It should be mentioned that, since the WLDR follows the LDA's process to obtain the transformation matrix after estimating  $\mathbf{S}_B$  and  $\mathbf{S}_W$ , if the worst-case scenario does not exist in the data set, WLDR is likely to yield a limited improvement, as evidenced by results in Figs. 1–3. Moreover, although we employ Loog's dissimilarity measure in this paper, other statistical measures can also be employed in the WLDR. But no measure can guarantee to yield the best results for all data sets, as can be seen in columns 4 and 5 in Tables 2 and 3 that the WLDR cannot outperform the aPAC always.

After presenting WLDR, we employ the evolution strategies to search for optimal weighting vectors with respect to

classification accuracy on the validation data sets. Since the purpose of employing the ES is to search for the optimal parameters, the ES can also be applied with the aPAC, F-LDA and other variants. EWLDR can also be viewed as another version of F-LDA. In F-LDA, the weights of between-class covariance matrices are manually selected in every iteration, while in the EWLDR algorithm, the weights are obtained by the ES procedure, and are incorporated in the estimation of both the between-class and the within-class covariance matrices. The experimental results demonstrate that the weights obtained from statistical approaches are usually sub-optimal and the EWLDR can yield a distinct improvement over them.

## 6. Summary

The linear discriminant analysis (LDA) is one of the most traditional linear dimensionality reduction methods. By transforming a data set into a new space according to Fisher's criterion, it preserves most of the discriminative information contained in the original data set. However, it only provides a sub-optimal solution in many situations with respect to real world problems. In view of this, this paper incorporates the inter-class relationships as relevance weights into the estimation of the overall within-class scatter matrix in order to improve the performance of the basic LDA method and some of its improved variants. In order to optimize the relevance weighting vectors, we employ the evolution strategy in our relevance-weighting framework. The experimental results have shown that in the new space achieved by our algorithm, higher classification accuracy can be achieved, implying that our algorithm can successfully preserve more discriminative information. Moreover, our optimization framework can also be applied to several existing improved versions of the LDA.

## References

- [1] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, New York, 1990.
- [2] Z. Jin, J.Y. Yang, Z.S. Hu, Z. Lou, Face recognition based on the uncorrelated discriminant transformation, Pattern Recognition 34 (2001) 1405–1416.
- [3] I. El-Feghi, M.A. Sid-Ahmed, M. Ahmadi, Automatic localization of craniofacial landmarks for assisted cephalometry, Pattern Recognition 37 (2004) 609–621.
- [4] S. Petridis, S.J. Perantonis, On the relation between discriminant analysis and mutual information for supervised linear feature extraction, Pattern Recognition 37 (2004) 857–874.
- [5] N. Kumar, A.G. Andreou, Heteroscedastic discriminant analysis and reduced rank HMMS for improved speech recognition, Speech Comm. 26 (1998) 283–297.
- [6] M. Loog, R.P.W. Duin, Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2004) 732–739.
- [7] T. Hastie, R. Tibshirani, Discriminant analysis by Gaussian mixtures, J. Roy. Statist. Soc. B 58 (1996) 155–176.
- [8] J.P. Hoffbeck, D.A. Landgrebe, Covariance matrix estimation and classification with limited training data, IEEE Trans. Pattern Anal. Mach. Intell. 18 (1996) 763–767.
- [9] J.H. Friedman, Regularized discriminant analysis, J. Am. Statist. Assoc. 84 (1989) 165–175.
- [10] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional data—with application to face recognition, Pattern Recognition 34 (2001) 2067–2070.
- [11] C. Liu, H. Wechsler, Robust coding schemes for indexing and retrieval from large face databases, IEEE Trans. on image processing 9 (2000) 132–137.
- [12] M. Loog, R.P.W. Duin, R. Haeb-Umbach, Multiclass linear dimension reduction by weighted pairwise fisher criteria, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 762–766.
- [13] R. Lotlikar, R. Kothari, Fractional-step dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 623–627.
- [14] T. Back, H.P. Schwefel, An overview of evolutionary algorithms for parameter optimization, Evol. Comput. 1 (1993) 1–23.
- [15] D. Kim, S.-Y. Bang, A handwritten numeral character classification using tolerant rough set, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 923–937.
- [16] C. Liu, H. Wechsler, Evolutionary pursuit and its application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 570–582.
- [17] M.L. Raymer, W.F. Punch, E.D. Goodman, L.A. Kuhn, A.K. Jain, Dimensionality reduction using genetic algorithms, IEEE Trans. on Evol. Comput. 4 (2000) 164–171.
- [18] M.M. Rizki, M.A. Zmuda, L.A. Tamburino, Evolving pattern recognition systems, IEEE Trans. Evol. Comput. 6 (2002) 594–609.
- [19] M. Kotani, M. Nakai, K. Akazawa, Feature extraction using evolutionary computation, Proceedings of the 1999 Congress on Evolutionary Computation, 1999.
- [20] Y. Liu, X. Yao, T. Higuchi, Evolutionary ensembles with negative correlation learning, IEEE Trans. Evol. Comput. 4 (2000) 380–387.
- [21] K.G. Khoo, P.N. Suganthan, Structural pattern recognition using genetic algorithms with specialized operators, IEEE Trans. Systems Man Cybernet. Part B 33 (1) (2003) 156–165.
- [22] C.L. Blake, C.J. Merz, UCI Repository of Machine Learning Databases, University of California, Irvine, Department of Information and Computer Sciences, 1996, <http://www.ics.uci.edu/~mllearn/MLRepository.html>.
- [23] R.A. Fisher, The statistical utilization of multiple measurements, Ann. Eugenics 8 (1938) 376–386.
- [24] C.R. Rao, The utilization of multiple measurements in problems of biological classification, J. Roy. Statist. Soc. B 10 (1948) 159–203.
- [25] X. Yao, Y. Liu, Fast evolution strategies, Control Cybernet. 26 (1997) 467–496.
- [26] M.M. Islam, X. Yao, K. Murase, A constructive algorithm for training cooperative neural network ensembles, IEEE Trans. Neural Networks 14 (2003) 820–834.
- [27] H.-G. Beyer, K. Deb, On self-adaptive features in real-parameter evolutionary algorithms, IEEE Trans. Evol. Comput. 5 (2001) 250–270.

**About the Author**—E.K. TANG received his B.E. Degree from Department of Control Science and Engineering of Huazhong University of Science and Technology, Wuhan, P. R. China in 2002. He is currently a Ph.D. student in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include pattern recognition, machine learning and evolutionary algorithms.

**About the Author**—PONNUTHURAI NAGARATNAM SUGANTHAN received the B.A degree, Postgraduate Certificate and M.A degree in Electrical and Information Engineering from the University of Cambridge, UK in 1990, 1992 and 1994, respectively. He obtained his Ph.D. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He was a predoctoral Research Assistant in the Department of Electrical Engineering, University of Sydney in 1995-96 and a lecturer in the Department of Computer Science and Electrical Engineering, University of Queensland in 1996-99. Since 1999 he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore where he was an Assistant Professor and now is an Associate Professor. He is an associate editor of the Pattern Recognition Journal and International Journal of Computational Intelligence. His research interests include evolutionary computation, applications of evolutionary computation, pattern recognition, bioinformatics and neural networks. He is a senior member of the IEEE and an associate member of the IEE.

**About the Author**—Professor XIN YAO obtained his BSc in 1982, MSc in 1985 and Ph.D. in 1990, all in computer science. He joined the University of Birmingham from Australia as a professor of computer science in 1999. He is a fellow of IEEE, the editor-in-chief of IEEE Transactions on Evolutionary Computation, an associate editor of several other international journals, and the editor of the book series on “Advances in Natural Computation” from World Scientific Publishing Co. He has been an invited keynote or plenary speaker of 30 international conferences in 11 different countries and a chair/co-chair of 27 international conferences. He was an IEEE Computational Intelligence Society Distinguished Lecturer in 2003. He won the prestigious IEEE Donald G. Fink Prize Paper Award (2001). He is currently the Director of The Centre of Excellence for Research in Computational Intelligence and Applications (CERCIA) at the University of Birmingham. He is also a Distinguished Visiting Professor at the University of Science and Technology of China and a visiting professor of three other universities. He has more than 200 research publications, including 63 refereed journal papers. His major research interests include evolutionary computation, neural network ensembles, global optimisation, evolvable hardware, data mining, and computational time complexity of evolutionary algorithms.

**About the Author**—A.K. QIN received his B.E. Degree from Department of Automatic Control Engineering of Southeast University, Nanjing, P. R. China in 2001. He is currently working towards the Ph.D. degree in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include machine learning, pattern recognition, neural network, evolutionary algorithms and bioinformatics. He is a member of Pattern Recognition and Machine Intelligence Association, Singapore.