# A Novel Nonparametric Linear Discriminant Analysis for High-Dimensional Data Classification

Jinn-Min Yang[1], Pao-Ta Yu[2]

*Department of Computer Science and Information Engineering, National Chung Cheng University*
*168 University Rd., Min-Hsiung 621, Chia-Yi, Taiwan, R.O.C.*

[1]ygm@ms3.ntcu.edu.tw
[2]csipty@cs.ccu.edu.tw

***Abstract*— Linear discriminant analysis (LDA) has played an important role for dimension reduction in patter recognition field. Basically, LDA has three deficiencies in dealing with classification problems. First, LDA is well-suited only for normally distributed data. Second, the number of features can be extracted are limited by the rank of between-class scatter matrix. Third, the singularity problem arises when dealing with high-dimensional data with small training samples. Nonparametric dimension reduction method, such as nonparametric discriminant analysis (NDA), is able to overcome the first two limitations of LDA, and the last problem is often resolved by regularization or eigen-decomposition techniques. In this study, we propose a novel nonparametric dimension reduction method with regularization to overcome all of the previously mentioned problems. The experimental results on real datasets demonstrate that the proposing method obtains satisfactory results, even in badly posed classification conditions.**

***Keywords*—Linear discriminant analysis (LDA), dimension reduction, feature extraction, small sample size problem, regularization.**

## 1. INTRODUCTION

Linear discriminant analysis (LDA) [1] has been played an important role for data classification. It is one of the most well-known dimension reduction methods and has been successfully applied to many classification problems. The purpose of LDA is to find a linear transformation that can be used to project data from a high-dimensional space into a low-dimensional subspace to mitigate the curse of dimensionality [2], [3] or the Hughes phenomenon [4], [5]. the Hughes phenomenon states that the ratio of the number of training samples and the number of features must be maintained at or above some minimum value to achieve statistical confidence [5]. However, it is not necessary to have sufficient training samples to keep the ratio in a high-dimensional classification task. Thus, by feature extraction, the ratio can be relatively enlarged. The curse of dimensionality can therefore be improved and the computational time can be reduced as well.

Basically, LDA has three inherent deficiencies in dealing with classification problems. First, LDA is only well-suited for normally distributed data [1]. If the distributions are significantly non-normal, the use of LDA cannot be expected to accurately indicate which features should be extracted to preserve complex structures needed for classification. Second, since the rank of between-class scatter matrix is the number of classes ($L$) minus one [1], the number of features can be extracted at most remains the same. Third, the singularity problem arises when dealing with high-dimensional and small sample size (SSS) data [1], [3], [6], [7]. Generally, there are three categories for solving the singularity of within-class scatter matrix [8]. In recent years, many approaches have been proposed to deal with the singularity problem for different applications, including PCA+LDA [9], regularized LDA (RLDA) [10], LDA/GSVD [11], LDA/QR [12], nonparametric weighted feature extraction (NWFE) [6] and fuzzy linear feature extraction (FLFE) [7]. Regularization and eigen-decomposition are the most often used techniques. However, the first two problems still exist. Nonparametric linear discriminant analysis such as nonparametric discriminant analysis (NDA) [13] provides a solution for circumventing both of the previously mentioned problems. In NWFE, a regularization technique is employed to solve

the singularity problem, and all problems of LDA are then resolved. Additionally, nonparametric feature extraction is generally of full rank which provides the ability to specify the number of extracted features desired and works well even for non-normally distributed data [6].

In this paper, a novel nonparametric dimension reduction method, called nonparametric linear discriminant analysis (NLDA), is proposed. Two techniques, regularization and feature adjustment [1, p.32], are adopted to circumvent the singularity problem and enhance the classification performance, respectively. The effectiveness of the proposed NLDA is evaluated by two hyperspectral datasets with different training sample sizes, including the ill-posed and poorly posed classification problems [14].

The rest of the paper is organized as follows. In Section 2, some related dimension reduction methods are reviewed. Then the details of the proposing nonparametric feature extractions are described in Section 3, followed by experimental designs and results in Section 4. Finally, conclusions are drawn in Section 5.

## 2. REVIEWS OF SOME DIMENSION REDUCTION METHODS

In this section, some work related to ours is reviewed. For convenience, some important notations employed in the study are presented in Table 1.

**TABLE 1.**
**SOME IMPORTANT NOTATIONS EMPLOYED IN THE PAPER**

| Notation | Description | Notation | Description |
|---|---|---|---|
| $X$ | data matrix | $N$ | total training samples |
| $X_i$ | data matrix of the $i$th class | $N_i$ | number of training samples in the $i$th class |
| $m_i$ | mean of the $i$th class | $x_\ell^{(i)}$ | the $\ell$th sample in the $i$th class |
| $m$ | global mean | $p$ | dimensionality of the reduced subspace |
| $L$ | number of classes | $P_i$ | prior probability of the $i$th class |
| $d$ | dimensionality of the original space | $M_j(x_\ell^{(i)})$ | local mean of $x_\ell^{(i)}$ in the $j$th class |
| $A$ | transformation matrix | $\mathcal{M}_j(x_\ell^{(i)})$ | weighted mean of $x_\ell^{(i)}$ in the $j$th class in NWFE |

### 2.1. Linear discriminant analysis (LDA)

In LDA [1], three scatter matrices, namely between-class, within-class and mixture scatter matrices, are defined as follows:

$$S_b = \frac{1}{N}\sum_{i=1}^{L} N_i (m_i - m)(m_i - m)^T, \quad (1)$$

$$S_w = \frac{1}{N}\sum_{i=1}^{L}\sum_{\ell=1}^{N_i} (x_\ell^{(i)} - m_i)(x_\ell^{(i)} - m_i)^T, \quad (2)$$

$$S_t = \frac{1}{N}\sum_{i=1}^{L}\sum_{\ell=1}^{N_i} (x_\ell^{(i)} - m)(x_\ell^{(i)} - m)^T, \quad (3)$$

where $L$ is the number of classes, $m_i$ and $m$ represent the $i$th class mean and the grand mean, respectively.

The goal of LDA is to find a transformation matrix $A$ which maximizes between-class and minimizes the within-class scatter matrices in the reduced dimensional space [1], [11]. The common optimization criterion for finding $A$ is

$$A = \arg\max_A \operatorname{tr}((A^T S_2 A)^{-1} A^T S_1 A). \quad (4)$$

where $(S_1, S_2) = (S_b, S_w)$ or $(S_1, S_2) = (S_b, S_t)$. The maximization of (4) is equivalent to solving the generalized eigenvalue decomposition problem

$$S_b v_h = \lambda_h S_w v_h, \quad h = 1, \ldots, p, \quad p \le L - 1. \quad (5)$$

where $p$ denotes the dimensionality of the reduced subspace, $(\lambda_h, v_h)$ represent the eigen-pair of $S_w^{-1} S_b$, and $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p$. Thus, the transformation matrix $A = [v_1, \ldots, v_p]$ can be obtained.

### 2.2. Regularized Linear Discriminant Analysis (RLDA)

In the regularized LDA (RLDA) [10], when $S_w$ (or $S_t$) is singular or ill-conditioned, a constant $\alpha$ is added to the diagonal elements of $S_w$ by

$$S_w^R = S_w + \alpha I_d \ (\text{or } S_t^R = S_t + \alpha I_d), \quad (6)$$

where $\alpha \in [0, \infty)$ and $I_d$ is the identity matrix of size $d$. It is easy to verify that (6) is positive definite [15]. Thus, the transformation matrix $A = [v_1, \ldots, v_p]$ consists of the eigenvectors of $(S_w^R)^{-1} S_b$ (or $(S_t^R)^{-1} S_b$), where $p \le L - 1$.

### 2.3. Nonparametric Discriminant Analysis (NDA)

LDA is called a parametric feature extraction method in [1], since it uses the mean vector and covariance matrix of each class. Fukunaga and Mantock [13] proposed a linear discriminant analysis with a nonparametric

between-class scatter matrix, called nonparametric discriminant analysis (NDA), overcoming the limitation for only $L-1$ features can be extracted at most.

The nonparametric between-class scatter matrix for the two-class problem in NDA, $S_b^{NDA}$, is represented as

$$S_b^{NDA} = \frac{1}{N}\sum_{\ell=1}^{N_1} w_\ell \left(x_\ell^{(1)} - M_2(x_\ell^{(1)})\right)\left(x_\ell^{(1)} - M_2(x_\ell^{(1)})\right)^T + \frac{1}{N}\sum_{\ell=1}^{N_2} w_\ell \left(x_\ell^{(2)} - M_1(x_\ell^{(2)})\right)\left(x_\ell^{(2)} - M_1(x_\ell^{(2)})\right)^T, \quad (7)$$

where $N_1 + N_2 = N$ and

$$M_j(x_\ell^{(i)}) = \frac{1}{k}\sum_{s=1}^{k} x_{sNN}^{(j)}, \quad (8)$$

denotes the sample mean of the $k$ NN's with respect to $x_\ell^{(i)}$ and is called the local mean of $x_\ell^{(i)}$ in class $j$. The weighting function $w_\ell$ is defined as

$$w_\ell = \frac{\min\left\{\mathrm{d}^\alpha\left(x_\ell, x_{kNN}^{(1)}\right), \mathrm{d}^\alpha\left(x_\ell, x_{kNN}^{(2)}\right)\right\}}{\mathrm{d}^\alpha\left(x_\ell, x_{kNN}^{(1)}\right) + \mathrm{d}^\alpha\left(x_\ell, x_{kNN}^{(2)}\right)}, \quad (9)$$

where $\alpha$ is a control parameter between zero and infinity, and $\mathrm{d}(x_\ell, x_{kNN}^{(i)})$ is the distance from $x_\ell$ to its $k$NN in class $i$.

The weighting function (9) is capable of achieving the goal of emphasizing the importance of boundary points. It takes on values close to 0.5 and drops off to zero as we move away from the class boundary. Obviously, the weighting function has the property that samples near the class boundary are given higher weights and those far away from the class boundary are given less. Nevertheless, if we focus on those samples near the class boundary, some problems occur. For example, if $x_\ell$ is a sample in class 1 and $\mathrm{d}^\alpha(x_\ell, x_{kNN}^{(2)})$ is small, then $x_\ell$ is considered to be more close to the class boundary but gains less weight. We can imagine that the weighting mechanism will notably affect the performance of NDA, especially when some classes are highly overlapped. In addition, the within-class scatter matrix of NDA is the same as LDA, so NDA still suffers from the singularity problem when the training sample size is small.

## 2.4. Nonparametric Weighted Feature Extraction (NWFE)

The main ideas of nonparametric weighted feature extraction (NWFE) [6] are placing different weights on every sample to compute the "weighted means", and applying the distances between samples, their weighted means as their "closeness" to boundary. Additionally, NWFE addressed a regularized within-class scatter matrix for alleviating the singularity. As a result, NWFE prevents the disadvantages of LDA and NDA and obtains satisfactory results [6]. The between-class scatter matrix $S_b^{NW}$ and the within-class scatter matrix $S_w^{NW}$ of NWFE are defined as

$$S_b^{NW} = \sum_{i=1}^{L} P_i \sum_{\substack{j=1,\\ j\neq i}}^{L}\sum_{\ell=1}^{N_i} w_\ell^{(i,j)} \left(x_\ell^{(i)} - \mathcal{M}_j(x_\ell^{(i)})\right)\left(x_\ell^{(i)} - \mathcal{M}_j(x_\ell^{(i)})\right)^T, \quad (10)$$

$$S_w^{NW} = \sum_{i=1}^{L} P_i \sum_{\ell=1}^{N_i} w_\ell^{(i,i)} \left(x_\ell^{(i)} - \mathcal{M}_i(x_\ell^{(i)})\right)\left(x_\ell^{(i)} - \mathcal{M}_i(x_\ell^{(i)})\right)^T, \quad (11)$$

where the scatter matrix weight $w_\ell^{(i,j)}$ is defined by

$$w_\ell^{(i,j)} = \frac{\mathrm{d}(x_\ell^{(i)}, \mathcal{M}_j(x_\ell^{(i)}))^{-1}}{\sum_{t=1}^{N_i}\mathrm{d}(x_t^{(i)}, \mathcal{M}_j(x_t^{(i)}))^{-1}}, \quad (12)$$

and the weighted mean is

$$\mathcal{M}_j(x_\ell^{(i)}) = \sum_{t=1}^{N_j} \eta_{\ell t}^{(i,j)} x_t^{(j)}, \quad (13)$$

and

$$\eta_{\ell t}^{(i,j)} = \frac{\mathrm{d}(x_\ell^{(i)}, x_t^{(j)})^{-1}}{\sum_{t=1}^{N_j}\mathrm{d}(x_\ell^{(i)}, x_t^{(j)})^{-1}}. \quad (14)$$

Equations (13) and (14) show that each sample $x_\ell^{(i)}$ has its own weighted mean $\mathcal{M}_j(x_\ell^{(i)})$ which is contributed by each sample $x_t^{(j)}$ in class $j$ according to the distance between $x_\ell^{(i)}$ and $x_t^{(j)}$. The longer the distance between $x_t^{(j)}$ and $x_\ell^{(i)}$ is, the less the contribution of $x_t^{(j)}$ is. Then, the relationships between $x_\ell^{(i)}$ and $\mathcal{M}_j(x_\ell^{(i)})$ are employed to design $S_b^{NW}$, $S_w^{NW}$ and $w_\ell^{(i,j)}$, as demonstrated in (10), (11) and (12). Furthermore, in NWFE, the within-class scatter matrix $S_w^{NW}$ is regularized by

$$S_w^{RNW} = 0.5 S_w^{NW} + 0.5\mathrm{diag}(S_w^{NW}). \quad (15)$$

Evidently, this regularization form reduces the values of the off-diagonal entries of $S_w^{NW}$ to half and keeps the diagonal entries invariant. The main disadvantage of NWFE is that it needs more computational time on computing the weighted mean of each class of each sample, particularly when the size of training samples is large.

## 3. NONPARAMETRIC LINEAR DISCRIMINANT ANALYSIS (NLDA)

In this section, the proposing algorithm, called nonparametric linear discriminant analysis (NLDA), will be introduced. The within-class scatter matrix of NLDA (denoted as $S_w^G$) is defined as

$$S_w^G = \sum_{i=1}^{L} P_i \sum_{\ell=1}^{N_i} (x_\ell^{(i)} - M_i)(x_\ell^{(i)} - M_i)^T, \quad (16)$$

where $P_i$ and $M_i$ are the prior probability and local mean with respect to $x_\ell^{(i)}$ in class $i$, respectively.

The between-class scatter matrix of NLDA ($S_b^G$) is defined as

$$S_b^G = \sum_{i=1}^{L} P_i \sum_{\substack{j=1 \\ j\neq i}}^{L} \sum_{\ell=1}^{N_i} (x_\ell^{(i)} - M_j)(x_\ell^{(i)} - M_j)^T, \quad (17)$$

where $M_j$ is the local mean with respect to $x_\ell^{(i)}$ in class $j$.

The idea to construct NLDA is twofold: First, we find that the local mean $M_j(x_\ell^{(i)})$ can be regarded as a leave-($N_i - k$)-out mean vector. Intuitively, $M_j(x_\ell^{(i)})$ can approximate to class mean $m_j$ as the value of $k$ is close to $N_j$. The estimators of scatter matrices will be more general and flexible. Second, to work well for non-normally distributed data, the within-class and between-class scatter matrices should be nonparametric simultaneously. The geometric depiction of the relationships of the within-class and between-class scatter matrices for the proposed NLDA is demonstrated in Fig. 1. The orange and green dash lines show the relationships between local means and class means in within-class and between-class, respectively.

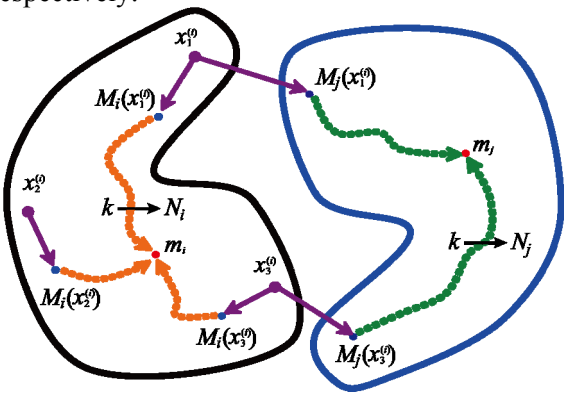



Fig. 1. Geometric depiction on the relationships between the local and class means.

### 3.1. Regularization

For extracting informative features, the criterion $J = tr(S_w^{-1} S_b)$ requires the within-class scatter matrix $S_w$ to be nonsingular [1], [11]. However, when the size of training samples is small, $S_w$ is often singular or nearly singular. For preventing the singularity of $S_w$, the regularization is one of the prominent techniques [6], [7], [10]. For NLDA, the form (15) employed in NWFE is taken. Because the selection of regularization parameter $\alpha$ in (18) is a model-selection problem [16], instead of using 0.5 as the regularization parameter, we adopt

$$S_w^{RG} = (1-\alpha) S_w^G + \alpha \mathrm{diag}(S_w^G). \quad (18)$$

The grid-search and cross validation (CV) methods are adopted to search the best value of $\alpha$ in this study.

### 3.2. Feature Adjustment

Simultaneous diagonalization of two matrices is a very powerful tool in pattern recognition [1]. In fact, the transformation matrix $A$ consists of eigenvectors of $(S_w^{RG})^{-1} S_b^G$ can diagonalize $S_w^{RG}$ and $S_b^G$ simultaneously, which has been proven in [1, p.32]. Nevertheless, when the singularity problem of $S_w^G$ has resolved by utilizing $S_w^{RG}$, there is another essential issue about the eigenvectors has to be taken care. That is, the matrix $(S_w^{RG})^{-1} S_b^G$ may be not symmetric in general, and subsequently the eigenvectors $v_h$'s are not mutually orthogonal. Thus, to make the $v_h$'s orthonormal with respect to $S_w^{RG}$ to satisfy $A^T S_w^{RG} A = I$, the scale of $v_h$ must be adjusted by

$$v_h = \frac{v_h}{\sqrt{v_h^T S_w^{RG} v_h}} \quad (19)$$

such that

$$\frac{v_h^T}{\sqrt{v_h^T S_w^{RG} v_h}} S_w^{RG} \frac{v_h}{\sqrt{v_h^T S_w^{RG} v_h}} = 1. \quad (20)$$

The proposed algorithm is described as follows.

**Algorithm :** NLDA

**Input**: the data matrix $X \in R^{d \times N}$, where $d$ is the dimensionality of the original space and $N$ is the number of training samples.

**Output**: the projection data matrix $Y = A^T X \in R^{p \times N}$, where $A \in R^{d \times p}$ and $p$ is the dimensionality of the reduced subspace.

Step 1. Select a value of $k$ for estimating the local means, $M_i(x_\ell^{(i)})$ and $M_j(x_\ell^{(i)})$, with respect to each training sample $x_\ell^{(i)}$ in $X$.

Step 2. Compute the within-class and between-class scatter matrices in (16) and (17), respectively.

Step 3. Calculate the regularized within-class scatter matrix $S_w^{RG}$ in (18).

Step 4. Select the $p$ eigenvectors of $(S_w^{RG})^{-1} S_b^G$, which correspond to the $p$ largest eigenvalues.

Step 5. Adjust each eigenvector $v_h$ by (19), $h = 1, \ldots, p$, and $A = [v_1, \ldots, v_p] \in R^{d \times p}$.

Step 6. Calculate the transformed data $Y = A^T X$.

## 4. Experimental Design and Results

### 4.1. Data Set

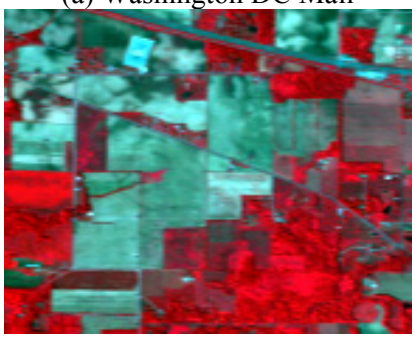
For evaluating the performance of the proposed NLDA, two real hyperspectral image datasets, the Washington DC Mall (DC) [4] and Indian Pine Site (IPS) [4], are employed. The DC and IPS datasets are an urban site and forest/agricultural site, respectively. They were gathered by a sensor known as the AVIRIS (Airborne Visible/Infrared Imaging Spectrometer). There are seven information classes in the DC dataset and eight classes in the IPS dataset. The dimensionality of DC and IPS datasets is 191 and 220, respectively.

### 4.2. Experimental Design

A portion of the original DC image and IPS image are selected as a test field, as shown in Fig. 2. Two different cases, each class with 20 (case I: $N_i = 20 < N < d$ ) and 40 (case II: $N_i = 40 < d < N$ ) training samples are investigated to discover the effect on the sizes of training samples in the experiments. In both cases, the test sample size of each class is 100. The cases I and II are the so-called ill-posed and poorly posed classification problems [7], [14], respectively. They are challenging cases in the field of pattern recognition. In each case, the training and testing datasets are randomly selected, and the experiment will be repeated 10 times. The average classification accuracies and their corresponding standard deviations by using 1 to 15 features will be computed.

Two linear feature extraction methods, NWFE and LDA, are used to compare the classification performance with the proposed NLDA. In this study, the 1-nearest neighbour (1NN) and soft-margin SVM with RBF kernel function (SVM-RBF) classifiers are used, which are implemented in PRTools [17] and LIBSVM [18], respectively. For the soft-margin SVM classifier, there is a parameter $C$ to control the trade-off between the margin and the size of the slack variables, and a parameter $\sigma$ for the RBF kernel function. We use the five-fold cross validation to find the best $C$ and $\sigma$ within the given set $\{10^{-5}, \ldots, 10^{5}\}$. The values of $k$ for estimating the local mean in NLDA are selected from {3, 5, 7} by using five-fold cross validation as well.

(a) Washington DC Mall



(b) Indian Pine Site

Fig. 2 The color IR images of a portion of the Washington DC Mall and Indian Pine Site data are displayed in (a) and (b), respectively.

### 4.3. Experimental Results

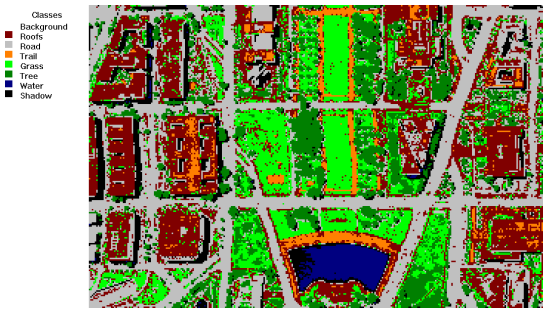
The best average classification accuracies (denoted as "acc") with their corresponding

standard deviations (denoted as "std") and number of features used (denoted as "$p$") are summarized in Table 2. Obviously, the proposed NLDA with 1NN classifier obtains the best results on the two datasets in all cases. The results also verify that the necessity of using more features than $L-1$ for classification. For example, NLDA with 1NN classifier on Washington DC Mall dataset with $N_i = 20$ obtains the best result when the dimensionality of the reduced subspace is 13, which is over $L-1$.

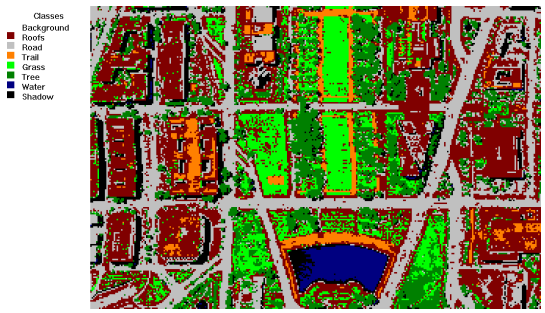
**TABLE 2**
**THE BEST AVERAGE CLASSIFICATION ACCURACIES (%) WITH THE CORRESPONDING STANDARD DEVIATIONS (%)AND THE NUMBER OF FEATURES USED FOR DIFFERENT FEATURE EXTRACTION METHODS BY USING 1NN AND SVM-RBF CLASSIFIERS.**

| Data Set | Feature Extraction | Classifier | $N_i = 20$ acc±std ($p$) | $N_i = 40$ acc±std ($p$) |
|---|---|---|---|---|
| DC | None | 1NN | 84.3±1.6 | 87.5±1.4 |
| | | SVM-RBF | 84.3±1.1 | 86.4±0.6 |
| | LDA | 1NN | 67.3±2.6 (6) | 88.2±1.9 (6) |
| | | SVM-RBF | 57.3±4.2 (6) | 82.3±2.3 (6) |
| | NLDA | 1NN | **90.8**±1.5 **(13)** | **93.1**±0.9 **(11)** |
| | | SVM-RBF | 90.3±1.1 (13) | 92.7±0.9 (13) |
| | NWFE | 1NN | 88±2.3 (5) | 91.2±1.2 (8) |
| | | SVM-RBF | 89.5±1.4 (4) | 91.2±0.9 (5) |
| IPS | None | 1NN | 66.7±1.0 | 71.9±1.0 |
| | | SVM-RBF | 68.9±2.6 | 70.2±1.9 |
| | LDA | 1NN | 62.2±2.2 (7) | 65.3±1.4 (7) |
| | | SVM-RBF | 62.5±2.2 (7) | 65.6±1.2 (7) |
| | NLDA | 1NN | **80.4**±1.4 **(8)** | **84.8**±1.0 **(7)** |
| | | SVM-RBF | 78.7±1.9 (8) | 82.4±1.5 (7) |
| | NWFE | 1NN | 79.8±2.1 (8) | 82.7±1.9 (7) |
| | | SVM-RBF | 78.2±2.1 (8) | 81.4±2.2 (14) |

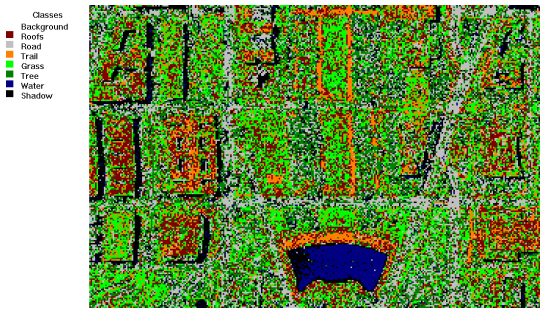
Some classified images of DC and IPC by using different feature extraction algorithms with 1NN classifier are demonstrated in Figs. 3 and 4, respectively. The DC thematic maps are obtained by applying linear feature extraction methods with 1NN classifier in case I ($N_i = 20$). Here $p$ is the number of features used by these methods with the highest accuracies in the corresponding methods. NLDA evidently obtains the best visual effect than the other methods since its excellent classification in "grass", "tree" and "roads" parts. For the performance on IPS dataset, NLDA outperforms other methods, particularly in "corn-notill" and "woods" parts. The variations in averaged classification accuracy obtained by applying NLDA and NWFE with 1NN classifier over different subspace dimensionality are analyzed with the help of plots shown in Fig. 5. For DC dataset, the accuracy of NWFE drops significantly as the dimensionality of the reduced subspace is more than 10. However, the accuracy of NLDA only varies little on DC dataset. On IPS dataset, there appears to be no appreciable difference between the performances of the two methods in $N_i = 20$ case. The classification accuracy of NLDA drops a little as the dimensionality of subspace is over 8, but NWFE does not. In Fig. 6, we demonstrate the variations in averaged classification accuracy obtained by applying NLDA with 1NN classifier over different $\alpha$ values on DC and IPS datasets in $N_i = 20$ case. The results show that there exists considerable performance difference between NLDA with and without ($\alpha = 0$) regularization technique. Also, the importance of introducing regularization for high-dimensional classification problem with small training samples is revealed.




(a) NLDA +1NN ($p = 13$)

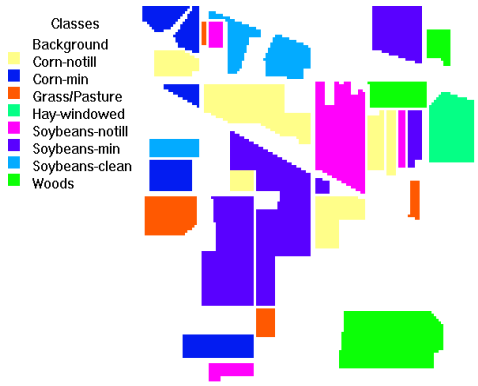



(b) NWFE +1NN ($p = 8$)

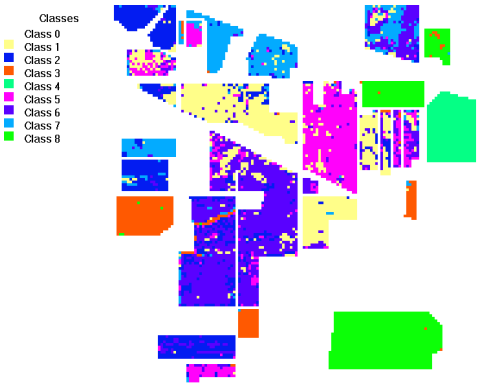



(c) LDA +1NN ($p = 6$)

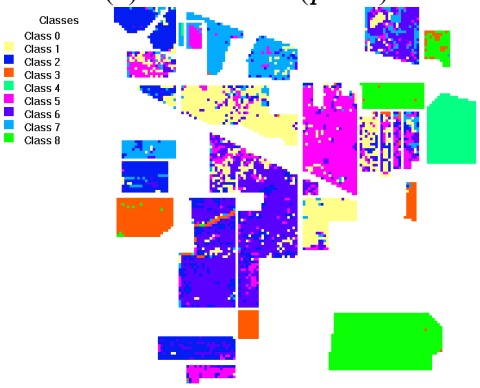
Fig. 3 Thematic maps resulting from the classification of the area of Fig. 2(a) in case I ($N_i = 20$). (a) to (c) are the results by NLDA, NWFE and LDA with 1NN classifier, respectively.



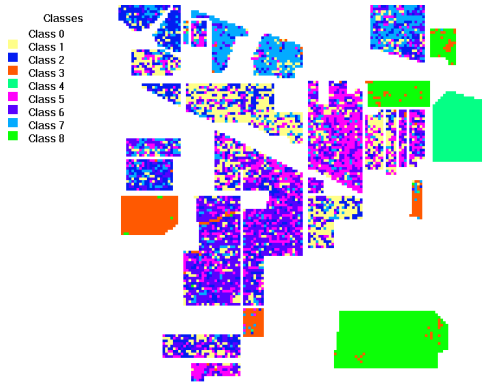

(a) The ground truth




(b) NLDA+1NN($p = 8$)

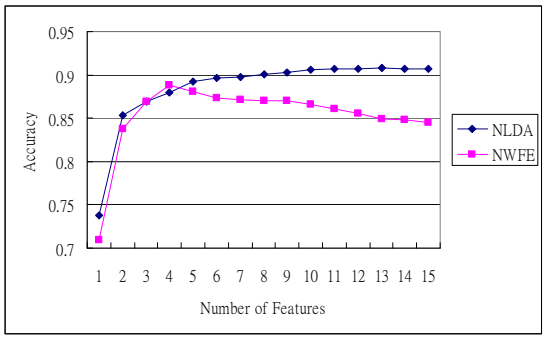



(c) NWFE+1NN ($p = 8$)

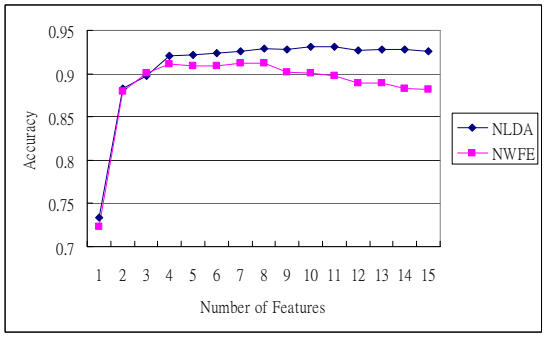



(d) LDA+1NN ($p = 7$)

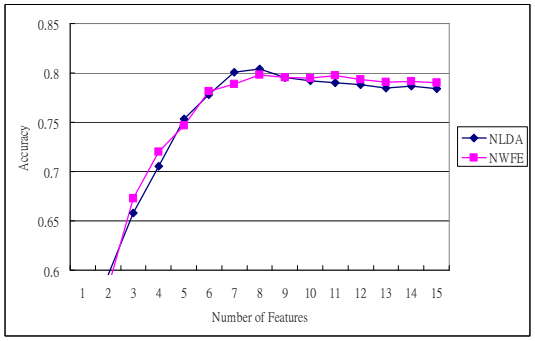
Fig. 4 Thematic maps resulting from the classification of the area of Fig. 2(b) in case II ($N_i = 40$). (a) is the ground truth of the area with eight classes, and (b) to (d) are the results applying NLDA, NWFE and LDA with 1NN classifier, respectively.



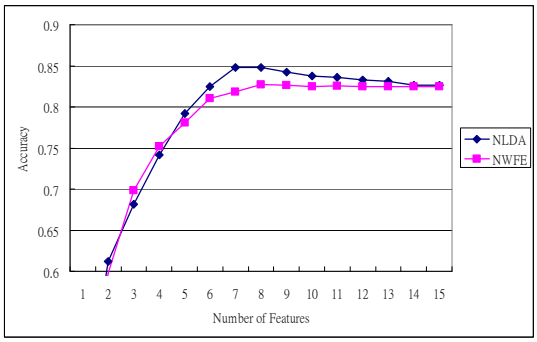

(a) DC ($N_i = 20$)




(b) DC ($N_i = 40$)

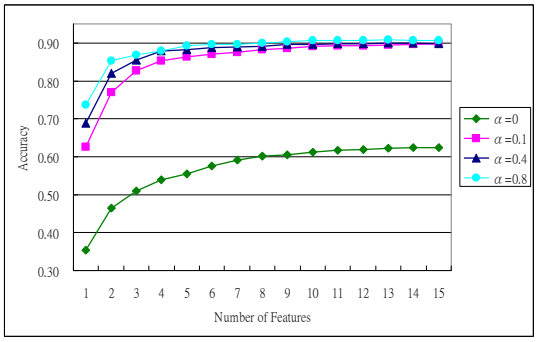



(c) IPS ($N_i = 20$)

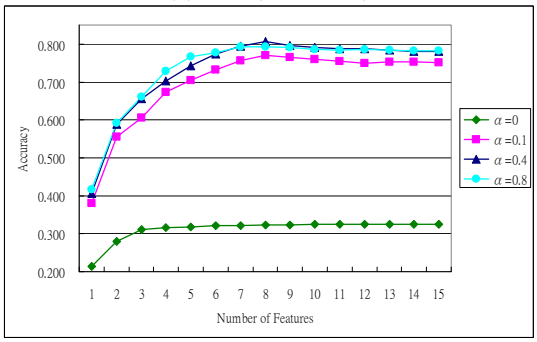



(d) IPS ($N_i = 40$)

Fig. 5. Variations in averaged classification accuracy obtained by applying NLDA and NWFE with 1NN classifier over different subspace dimensionality on Washington DC Mall and Indian Pine Site datasets.




(a) DC ($N_i = 20$)




(b) IPS ($N_i = 20$)

Fig. 6. The different averaged classification accuracy obtained by applying NLDA with 1NN classifier over different values of $\alpha$ on Washington DC Mall and Indian Pine Site datasets in case 1.

## 5. CONCLUSIONS

The objective of this study is to emphasize the necessities of developing a nonparametric feature extraction algorithm and relative issues for high-dimensional data classification when the size of the training samples is small. In this study, we propose a novel nonparametric linear discriminant analysis, NLDA, with two important techniques, regularization and feature adjustment, for improving its classification performance. As shown from the experimental results, NLDA has better classifiability than NWFE and LDA, even in the ill-posed and poorly posed classification situations. In addition, as supported from the results of NLDA and NWFE, the type of mean vector seems to be more important than the weighting parts for constructing a nonparametric feature extraction model.

## REFERENCES

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second ed., Academic Press, New York, 1990.

[2] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, second ed., John Wiley & Sons, New York, 2001.

[3] S.J. Raudys and A.K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol.13 no.3, pp. 252-264, 1991.

[4] D.A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*, John Wiley and Sons, Hoboken, Chichester, 2003.

[5] P.K. Varshney and M.K Arora. *Advanced image processing techniques for remotely sensed hyperspectral data*, Springer, New York, 2004.

[6] B.C. Kuo and D.A. Landgrebe, "Nonparametric weighted feature extraction for classification," *IEEE Transaction on Geoscience and Remote Sensing*, vol.42, no.5, pp.1096-1105, 2004.

[7] J.M. Yang, P.T. Yu, B.C. Kuo, T.Y. Hsieh, "A novel fuzzy linear feature extraction for hyperspectral image classification," in *Proc. of IEEE International Conference on Geoscience and Remote Sensing Symposium*, 2006, pp. 3895 - 3898.

[8] A. R. Webb, *Statistical Pattern Recognition*,

second ed., John Wiley and Sons, Hoboken, Chichester, 2002.

[9] P. N. Belhumeour, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 7, pp. 711–720, Jul. 1997.

[10] Y. Guo, T. Hastie, and R. Tibshirani, “Regularized linear discriminant analysis and its application in microarrays,” Biostatistics, vol. 8, no. 1, pp. 86–100, 2007.

[11] P. Howland, M. Jeon, and H. Park, “Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition,” SIAM J. Matrix Anal. Appl., vol. 25, no. 1, pp. 165–179, 2003.

[12] J. Ye and Q. Li, “LDA/QR: an efficient and effective dimension reduction algorithm and its theoretical foundation,” Pattern Recognition, vol.37, no.4, pp.851-854, 2004.

[13] K. Fukunaga and J.M. Mantock, “Nonparametric discriminant analysis,” IEEE Transaction on Pattern Analysis and Machine Intelligence, vol.5, pp. 671-678, 1983.

[14] A. Baraldi, L. Bruzzone, and P. Blonda, “Quality assessment of classification and cluster maps without ground truth knowledge,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 4, pp.857-873, April, 2005.

[15] G. H. Golub and C. F. Van Loan, *Matrix Computations,* Baltimore, MD: The Johns Hopkins University Press, 1996.

[16] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer-Verlag, 2001.

[17] R. P. W. Duin, “PRTools, a Matlab toolbox for pattern recognition,” [Online]. Available: http://www.prtools.org/, 2008.

[18] C. C. Chang and C. J. Lin, “LIBSVM : a library for support vector machines,” 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm.