

# Nonparametric Fuzzy Feature Extraction for Hyperspectral Image Classification

Jinn-Min Yang, Pao-Ta Yu, Bor-Chen Kuo, and Ming-Hsiang Su

## Abstract

Feature extraction plays an essential role in high-dimensional data classification. Linear discriminant analysis (LDA) is one of the most well-known methods for reducing data dimensionality in various fields. However, there are three inherent limitations when applying LDA to extract features. First, the number of features that can be extracted by LDA is the number of classes minus one at most. Second, it cannot perform well for non-normally distributed data. Third, it suffers from the singularity problem when handling the small sample size (SSS) problem. Nonparametric feature extraction algorithms such as nonparametric discriminant analysis (NDA) and nonparametric weighted feature extraction (NWFE) are developed to overcome the limitations of LDA and preserve better data structure in the reduced feature space for classification. In this study, we propose a novel nonparametric feature extraction method, called nonparametric fuzzy feature extraction (NFFE) method, to which some properties revealed from the fuzzification procedure of the fuzzy  $K$ -nearest neighbor algorithm are introduced. The performance of NFFE is investigated on two remotely sensed hyperspectral images with different training sample sizes, including the so-called ill-posed and poorly posed classification cases. The experimental results demonstrate that 1NN and SVM classifiers with NFFE features achieve better classification results than with features extracted from some existing methods.

**Keywords:** *Dimension reduction, Feature extraction, Hyperspectral image, Small sample size problem.*

## 1. Introduction

Feature extraction [1] is an important technique for high-dimensional classification problems, which aims to

mitigate the Hughes phenomenon [2], [3] or the curse of dimensionality [4] so as to enhance classification performance. The Hughes phenomenon describes the ratio of the training samples to the number of features that must be maintained at or above some minimum value to achieve statistical confidence [2]. However, it is not necessary to have sufficient training samples to keep the ratio in a high-dimensional classification task. Thus, by using feature extraction to reduce the data dimensionality, the ratio can be relatively enlarged without increasing the training samples. Feature extraction uses all the features to construct a transformation matrix that maps the original data to a low-dimensional subspace. As compared with feature selection [5], the main advantages of feature extraction are the information of all the original bands is used, and it is easier to use than feature selection for high-dimensional data [6].

Linear discriminant analysis (LDA) [5] is one of the well-known dimension reduction methods and has been successfully applied to many classification problems. LDA uses within-class, between-class and mixture (or total) scatter matrices to formulate criteria of class separability. 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 with maximized class separability [5], and the classification accuracy of the projected data can therefore be enhanced. However, LDA has three inherent deficiencies when dealing with classification problems. First, LDA is only well-suited for normal distributed data [5]. 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 minus one [5], the maximum number of features that can be extracted remains the same, which may not be sufficient for achieving better accuracy in practical application [7]. Third, the singularity problem arises when dealing with high-dimensional and small sample size (SSS) [8] problems. Lots of algorithms are proposed to overcome the drawbacks of LDA [7], [9]-[13].

LDA was called a parametric feature extraction method in [5] since it uses the mean vector and covariance matrix of each class, meaning that its theory is based on data with normal distribution. Fukunaga and

---

Corresponding Author: Jinn-Min Yang is with the Department of Mathematics Education, National Taichung University of Education, 140 Min-Shen Rd., Taipei, Taiwan, 403.

E-mail: jinnmin@gmail.com

Manuscript received 24 April. 2009; revised 29 April. 2010; accepted 14 August. 2010.

Mantock [9] proposed a discriminant analysis algorithm with a nonparametric between-class scatter matrix, namely nonparametric discriminant analysis (NDA). NDA aims to find a better between-class scatter matrix to overcome the limitation of LDA for only  $L - 1$  features can be extracted at most, and to project data as separate as possible. However, the within-class scatter matrix of NDA is the same as LDA, so NDA still suffers from the problem of singularity when the training sample size is small. Kuo and Landgrebe [7] addressed the nonparametric weighted feature extraction (NWFE) with fully nonparametric within- and between-class scatter matrices. In addition, it includes a regularization technique [14] to solve the singularity problem. Some studies [15]-[20] have shown that NWFE is efficient in reducing dimensionality. NWFE reveals some useful consequences for building a nonparametric feature extraction method.

In this paper, a novel nonparametric feature extraction method, called nonparametric fuzzy feature extraction (NFFE), is proposed. NFFE introduces fuzzy membership grades to design its within-class and between-class scatter matrices, and a more general regularization form with the same components in NWFE is adopted to alleviate the singularity problem. Importantly, a theoretical adjustment on features, not applying in LDA, NDA and NWFE, is taken to orthogonalize the features. The effectiveness of NFFE is evaluated on two remotely sensed hyperspectral image data sets. A hyperspectral image is generally with hundred of features and the cost of collecting its ground-truth can be considerably difficult and expensive. The SSS problem has been a key issue for hyperspectral image classification. Therefore, the feature extraction techniques have been played an important role in hyperspectral image classification.

The rest of the paper is organized as follows. In Section 2, some feature extraction methods are reviewed and the proposed NFFE is presented in Section 3, followed by the experimental designs and results in Section 4. Finally, conclusions are drawn in Section 5.

## 2. Related Work

### 2.1 Linear Discriminant Analysis

The goal of LDA is to find a transformation matrix  $A$  which maximizes the class separability [5]  $J = \text{tr}(S_w^{-1}S_b)$  in the transformed space, where  $S_w$  and  $S_b$  denote the within-class and between-class scatter matrices of LDA, respectively. That is

$$A = \underset{A}{\operatorname{argmax}} \operatorname{tr}((A^T S_w A)^{-1} A^T S_b A). \quad (1)$$

The  $S_w$  and  $S_b$  are defined as

$$S_b = \sum_{i=1}^L P_i (m_i - m_0)(m_i - m_0)^T, \quad (2)$$

$$S_w = \sum_{i=1}^L P_i S_i = S, \quad (3)$$

where  $P_i$  denotes the prior probability of class  $i$ ,  $L$  is the number of classes,  $m_0$  is the grand mean,  $S$  is the common covariance,  $m_i$  and  $S_i$  are the mean and covariance matrix of class  $i$ , respectively.

The maximization of (1) is equivalent to solving the generalized eigenvalue decomposition problem

$$S_b v_h = \lambda_h S_w v_h, \quad h = 1, \dots, p, \quad p \leq L - 1. \quad (4)$$

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

### 2.2 Nonparametric Discriminant Analysis

The nonparametric between-class scatter matrix  $S_b^{NDA}$  of NDA for the two-class problem is defined as

$$S_b^{NDA} = \frac{1}{N} \sum_{\ell=1}^{N_1} w_\ell \left( x_\ell^{(1)} - M_2(x_\ell^{(1)}) \right) \left( x_\ell^{(1)} - M_2(x_\ell^{(1)}) \right)^T + \frac{1}{N} \sum_{\ell=1}^{N_2} w_\ell \left( x_\ell^{(2)} - M_1(x_\ell^{(2)}) \right) \left( x_\ell^{(2)} - M_1(x_\ell^{(2)}) \right)^T, \quad (5)$$

where  $N_1 + N_2 = N$  and

$$M_j(x_\ell^{(i)}) = \frac{1}{k} \sum_{s=1}^k x_{sNN}^{(j)}, \quad (6)$$

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

$$w_\ell = \frac{\min \left\{ d^\alpha \left( x_\ell, x_{kNN}^{(1)} \right), d^\alpha \left( x_\ell, x_{kNN}^{(2)} \right) \right\}}{d^\alpha \left( x_\ell, x_{kNN}^{(1)} \right) + d^\alpha \left( x_\ell, x_{kNN}^{(2)} \right)}, \quad (7)$$

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

The weighting function (7) 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 classification boundary. The control parameter,  $\alpha$ , adjusts how rapidly  $w_\ell$  falls to zero as we move away. Obviously, the weighting function has the property that samples near the classification boundary are given higher values of the weights and those far away from the classification boundary are given less. Nevertheless, if we focus on those samples near the class boundary, an important problem will arise. For example, if  $x_\ell$  is a sample in class 1 and  $d^\alpha(x_\ell, x_{kNN}^{(2)})$  is small, then  $x_\ell$  is considered to be more close to the classification boundary but gains less weight. This phenomenon will happen in highly overlapped classes and notably affect the performance of NDA.

### 2.3 Nonparametric Weighted Feature Extraction

The main ideas of nonparametric weighted feature extraction (NWFE) [7] are to place different weights on every sample to compute the “weighted means”, and apply to the distances between samples and their weighted means as their “closeness” to classification boundary. Additionally, NWFE addresses 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 [7]. The between-class and within-class scatter matrices of NWFE, denoted as  $S_b^{NW}$  and  $S_w^{NW}$ , are defined as

$$S_b^{NW} = \sum_{i=1}^L P_i \sum_{\substack{j=1, \ell=1 \\ j \neq i}}^{N_i} \frac{w_{\ell}^{(i,j)}}{N_i} (x_{\ell}^{(i)} - M_j(x_{\ell}^{(i)})) (x_{\ell}^{(i)} - M_j(x_{\ell}^{(i)}))^T, \quad (8)$$

$$S_w^{NW} = \sum_{i=1}^L P_i \sum_{\ell=1}^{N_i} \frac{w_{\ell}^{(i,i)}}{N_i} (x_{\ell}^{(i)} - M_i(x_{\ell}^{(i)})) (x_{\ell}^{(i)} - M_i(x_{\ell}^{(i)}))^T, \quad (9)$$

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

$$w_{\ell}^{(i,j)} = \frac{d(x_{\ell}^{(i)}, M_j(x_{\ell}^{(i)}))^{-1}}{\sum_{t=1}^{N_i} d(x_t^{(i)}, M_j(x_t^{(i)}))^{-1}}, \quad (10)$$

and the weighted mean is

$$M_j(x_{\ell}^{(i)}) = \sum_{t=1}^{N_j} \eta_{\ell t}^{(i,j)} x_t^{(j)}, \quad (11)$$

and

$$\eta_{\ell t}^{(i,j)} = \frac{d(x_{\ell}^{(i)}, x_t^{(j)})^{-1}}{\sum_{t=1}^{N_j} d(x_{\ell}^{(i)}, x_t^{(j)})^{-1}}. \quad (12)$$

Equations (11) and (12) show that each sample  $x_{\ell}^{(i)}$  has its own weighted mean  $M_j(x_{\ell}^{(i)})$  in class  $j$ , which is contributed by each sample  $x_t^{(j)}$  in class  $j$  according to the distance between  $x_{\ell}^{(i)}$  and  $x_t^{(j)}$ . A longer distance between  $x_t^{(j)}$  and  $x_{\ell}^{(i)}$  implies less contribution of  $x_t^{(j)}$ . Then, the relationships between  $x_{\ell}^{(i)}$  and  $M_j(x_{\ell}^{(i)})$  are employed to design,  $S_b^{NW}$ ,  $S_w^{NW}$  and  $w_{\ell}^{(i,j)}$ , as demonstrated in (8), (9) and (10). Note that the criterion  $J = \text{tr}((S_w^{NW})^{-1} S_b^{NW})$  requires  $S_w^{NW}$  to be nonsingular for extracting sound features [5], [11]. However, when the size of training samples is small,  $S_w^{NW}$  is often singular or nearly singular. Regularization is employed to improve the singularity problem of  $S_w^{NW}$  in NWFE [7]. The Tikhonov regularization,  $(S_w^{CN})^{-1} = (S_w^{CN} + \alpha I)^{-1}$  where  $\alpha > 0$ , is not the best choice for the regularization in feature extraction and the diagonal part of the original matrix may be more important, as demonstrated from [21]-[23]. Instead of applying Tikhonov regularization,  $S_w^{NW}$  is replaced by the regularized within-class scatter matrix  $S_w^{NW_R}$

$$S_w^{NW_R} = 0.5 S_w^{NW} + 0.5 \text{diag}(S_w^{NW}). \quad (13)$$

Evidently,  $S_w^{NW_R}$  is constructed by reducing the ef-

fect of the cross products of  $S_w^{NW}$  to half and keeping the diagonal entries invariant. The main disadvantage of NWFE is that it takes enormous computation time on calculating the weighted mean and the weight of each sample in each class. The computation time of NWFE markedly increases with increasing  $N_i$ .

### 3. Nonparametric Fuzzy Feature Extraction

The ideas behind the nonparametric fuzzy feature extraction (NFFE) are described as follows. In NFFE, we use the membership grades estimated by the fuzzification procedure of the fuzzy  $K$ -nearest neighbor (FKNN) algorithm [24] to design its weighting function. The fuzzification procedure of FKNN is defined as

$$\mu_j(x_{\ell}^{(i)}) = \begin{cases} 0.51 + 0.49 \times \frac{n_j}{k_1} & \text{if } j = i \\ 0.49 \times \frac{n_j}{k_1} & \text{if } j \neq i \end{cases} \quad (14)$$

where  $x_{\ell}^{(i)}$  is the  $\ell$ th training sample in the class  $i$ ,  $k_1$  is a given constant denoting for the  $k_1$ -nearest neighbors of  $x_{\ell}^{(i)}$ ,  $n_j$  is the number of samples that belongs to the class  $j$  among the  $k_1$ -nearest neighbors of  $x_{\ell}^{(i)}$  and  $\sum_{j=1}^L n_j = k_1$ . Since  $x_{\ell}^{(i)}$  is labeled, the membership grade to which it belongs will be more than 0.51, and increases by the number of samples from the same class among the  $k_1$ -nearest neighbors. Therefore, each training sample  $x_{\ell}^{(i)}$  is associated with a membership vector  $\mu(x_{\ell}^{(i)}) = [\mu_1(x_{\ell}^{(i)}), \dots, \mu_L(x_{\ell}^{(i)})]$ . Here  $\mu_i(x_{\ell}^{(i)})$  denotes the membership grade of the class that  $x_{\ell}^{(i)}$  belongs to and is called the within-class membership grade in this study. The other membership grades in  $\mu(x_{\ell}^{(i)})$  is called between-class membership grades. If  $\mu_i(x_{\ell}^{(i)})$  is 1, then  $x_{\ell}^{(i)}$  is surrounded with  $k_1$  same-class nearest neighbors and is considered away from the class boundary, and a value of 0.51 represents  $x_{\ell}^{(i)}$  locates near the classification boundary or even an outlier. A two-class example demonstrating the computation of membership grades is illustrated in Figure 1, where  $k_1$  equals to 3. The membership vectors of some samples are labeled and those circled samples are regarded as the samples near the classification boundary. Thus, these samples should be emphasized by assigning larger weights. In fact, the fuzzification procedure devotes to collecting local information of each training sample, and the information exerts a considerable influence on the classification phase. The membership grades computed by (14) reflect the information, and we employ the information to design the weighting function of NFFE.

The within-class scatter matrix of NFFE ( $S_{fw}$ ) is defined as

$$S_{fw} = \sum_{i=1}^L P_i \sum_{\ell=1}^{N_i} a_{\ell}^{(i,i)} (x_{\ell}^{(i)} - \mathcal{M}_i(x_{\ell}^{(i)}))(x_{\ell}^{(i)} - \mathcal{M}_i(x_{\ell}^{(i)}))^T, \quad (15)$$

where  $\mathcal{M}_i(x_{\ell}^{(i)})$  is the weighted local mean of  $x_{\ell}^{(i)}$  in class  $i$ .  $a_{\ell}^{(i,i)}$  is the weight with respect to  $x_{\ell}^{(i)}$ . Here, we define the weighted local mean  $\mathcal{M}_i(x_{\ell}^{(i)})$  as

$$\mathcal{M}_i(x_{\ell}^{(i)}) = \sum_{s=1}^{k_2} \alpha_s^{(i)} x_{sNN}^{(i)}, \quad (16)$$

with  $\alpha_s^{(i)} = \frac{\mu_i(x_{sNN}^{(i)})}{\sum_{s=1}^{k_2} \mu_i(x_{sNN}^{(i)})}$ .  $x_{sNN}^{(i)}$  denotes the  $s$ th nearest neighbor of  $x_{\ell}^{(i)}$  in class  $i$ , and  $\mu_i(x_{sNN}^{(i)})$  represents  $x_{sNN}^{(i)}$ 's within-class membership grade.  $\mathcal{M}_i(x_{\ell}^{(i)})$  is computed by the  $k_2$ -nearest neighbors around  $x_{\ell}^{(i)}$ , and their corresponding membership grades as the weights. The weighting function  $a_{\ell}^{(i,i)}$  is defined as

$$a_{\ell}^{(i,i)} = 1 - (\mu_i(x_{\ell}^{(i)}) / \sum_{\ell=1}^{N_i} \mu_i(x_{\ell}^{(i)})). \quad (17)$$

As demonstrated from Figure 1, the within-class membership grades are small when samples are close to the classification boundary. Therefore, (17) can confirm that these samples are with larger weights.

The between-class scatter matrix of NFFE ( $S_{fb}$ ) is defined as

$$S_{fb} = \sum_{i=1}^L P_i \sum_{j=1}^L \sum_{\ell=1}^{N_i} b_{\ell}^{(i,j)} (x_{\ell}^{(i)} - \mathcal{M}_j(x_{\ell}^{(i)}))(x_{\ell}^{(i)} - \mathcal{M}_j(x_{\ell}^{(i)}))^T, \quad (18)$$

where  $\mathcal{M}_j(x_{\ell}^{(i)}) = \sum_{s=1}^{k_2} \alpha_s^{(j)} x_{sNN}^{(j)}$  and  $b_{\ell}^{(i,j)}$  are the weighted local mean and the weight of  $x_{\ell}^{(i)}$  in class  $j$ , respectively. Here  $b_{\ell}^{(i,j)}$  is defined as

$$b_{\ell}^{(i,j)} = \mu_j(x_{\ell}^{(i)}) / \sum_{\ell=1}^{N_i} \mu_j(x_{\ell}^{(i)}). \quad (19)$$

As shown from (14), the closer the sample  $x_{\ell}^{(i)}$  is to the  $j$ th class, the larger the between-class membership grade it will have. Therefore,  $b_{\ell}^{(i,j)}$  confirms that samples near classification boundary gain larger weights. In NFFE, the membership grades of each sample are applied to design the weighting functions and weighted local means.

The geometric depiction of the estimation of the within-class scatter matrices for LDA and the proposed NFFE is demonstrated in Figure 2. The plot (a) depicts the relationship between samples and class mean ( $m_i$ ) in LDA, and the plot (b) demonstrates the relationship between samples and weighted local mean ( $\mathcal{M}_i(x_{\ell}^{(i)})$ ) in NFFE. Obviously, the estimator of NFFE can describe the scatter of the data set better than that of LDA. Additionally, NFFE uses  $N_i$  samples to calculate the weighted mean  $\mathcal{M}_i(x_{\ell}^{(i)})$ , but NFFE employs only  $k_2$

samples to compute  $\mathcal{M}_i(x_{\ell}^{(i)})$ , so NFFE can reduce the computation time. The difference in computation time between NFFE and NFFE markedly increases with increasing  $N_i$ .

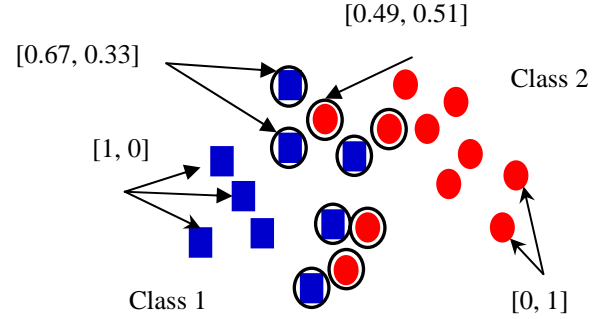


Figure 1. Illustration of the fuzzification procedure of FKNN algorithm with  $k_1 = 3$ .

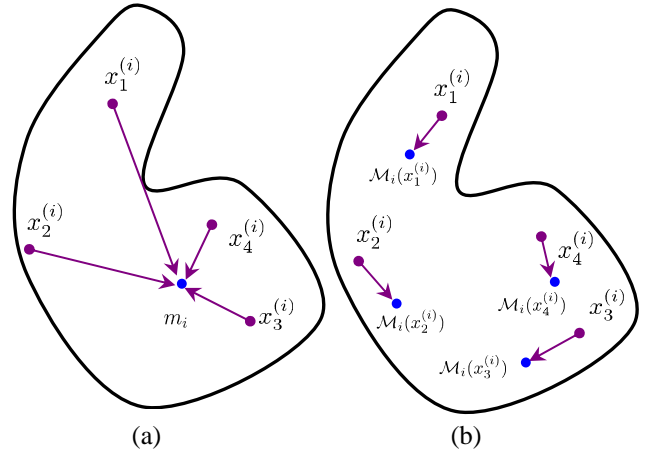


Figure 2. Geometric depiction on the estimation of within-class scatter matrices: (a)  $S_w$  in LDA; (b)  $S_{fw}$  in NFFE.

To make NFFE applicable for SSS problems, the regularization of  $S_{fw}$  is implemented by including an adaptive parameter for finding a suitable within-class estimator of NFFE. The regularized within-class scatter matrix  $S_{fw}^R$  of NFFE is defined as

$$S_{fw}^R = (1 - \mu)S_{fw} + \mu \text{diag}(S_{fw}), \quad 0 \leq \mu \leq 1. \quad (20)$$

where  $\mu$  is the regularization parameter. In NFFE,  $\mu$  is only fixed at 0.5. As is well-known, the selection of optimal values of parameters is a model-selection problem [25]. Thus, we take a more general form to find a more suitable within-class scatter matrix estimator for NFFE. In this study, the grid-search and cross validation (CV) methods are adopted to search the optimal value of  $\mu$ .

When the singularity problem of  $S_{fw}$  has resolved by utilizing  $S_{fw}^R$ , there exists another essential issue about the extracted eigenvectors. Simultaneous diagonalization of two matrices is a very powerful tool in pattern recog-

inition [5]. In fact, the transformation matrix  $A$  consists of eigenvectors of  $(S_{fw}^R)^{-1}S_{fb}$  can diagonalize  $S_{fw}^R$  and  $S_{fb}$  simultaneously, which has been proven in [5, p.32]. Nevertheless, the matrix  $(S_{fw}^R)^{-1}S_{fb}$  may be not symmetric in general, and subsequently the eigenvectors  $v_h$  are not mutually orthogonal. Instead, the  $v_h$ 's are orthogonal with respect to  $S_{fw}^R$ : that is,  $v_h^{-1}S_{fw}^R v_{h'} = 0$ , for  $h \neq h'$ . To make the  $v_h$ 's orthonormal with respect to  $S_{fw}^R$  to satisfy  $A^T S_{fw}^R A = I$ , the scale of  $v_h$  must be adjusted by

$$v_h = \frac{v_h}{\sqrt{v_h^T S_{fw}^R v_h}} \quad (21)$$

such that

$$\frac{v_h^T}{\sqrt{v_h^T S_{fw}^R v_h}} S_{fw} \frac{v_h}{\sqrt{v_h^T S_{fw}^R v_h}} = 1. \quad (22)$$

This feature adjustment procedure is theoretically important, but it is not included in LDA, NDA and NWF. We will make some efforts to investigate the effects of this procedure in this study.

As previously stated, NFFE includes three parts: the design of nonparametric scatter matrices, the regularization of the within-class scatter matrix, and the adjustment of the extracted features. The steps of the proposed NFFE are summarized in the following.

Algorithm: NFFE

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

Output: the transformed 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 transformed subspace.

- Step 1. Given a value of  $k_1$  for estimating the membership grades of samples in  $X$ .
- Step 2. Compute the within-class and between-class weights of each  $x_\ell^{(i)}$  in  $X$ , i.e.,  $a_\ell^{(i,i)}$  and  $b_\ell^{(i,j)}$ .
- Step 3. Given a value of  $k_2$  to compute the weighted local means and then compute  $S_{fw}$  and  $S_{fb}$  by (15) and (18), respectively.
- Step 4. Replace  $S_{fw}$  with  $S_{fw}^R$  (Eq. (20)) where the optimal value of  $\mu$  is searched by grid-search and 5-fold CV methods in a grid of 0.05.
- Step 5. Select the  $p$  eigenvectors of  $(S_{fw}^R)^{-1}S_{fb}$ , which correspond to the  $p$  largest eigenvalues.
- Step 6. Adjust each eigenvector  $v_h$  by (21), and  $A = [v_1, \dots, v_p] \in R^{d \times p}$
- Step 7. Calculate the transformed data  $Y = A^T X$ .

## 4. Experimental Design and Results

To evaluate the performance of the proposed NFFE, two remotely sensed hyperspectral image data sets are employed. The overall averaged accuracy with two classifiers, 1-nearest-neighbor (1NN) [4], [5] and support vector machine (SVM) [26] with RBF kernel function, on the transformed data will be reported.

### 4.1 Data Set

Two real data sets are applied to compare the performances of NFFE and other famous feature extraction methods, including the Indian Pines Site (IPS) image, a mixed forest/agricultural site in Indiana [1], and the Washington DC Mall (WDC) image [1] as an urban site. The IPS data set were gathered by a sensor known as the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) [27]. The WDC data set is a Hyperspectral Digital Imagery Collection Experiment (HYDICE) [28] airborne hyperspectral data flightline over the Washington, DC Mall.

The IPS image, mounted from an aircraft flown at 65000 ft. altitude and operated by the NASA/Jet Propulsion Laboratory, with the size of 145×145 pixels has 220 spectral bands measuring approximately 20m across on the ground. Since the size of samples in some classes are too small to retain enough disjoint samples for training and testing, only eight classes: Corn-notill, Corn-min, Grass/Pasture, Hay-windrowed, Soybeans-notill, Soybeans-min, Soybean-clean, and Woods, were selected for the experiments.

The Washington DC Mall (WDC) dataset is available in the student CD-ROM of [1]. Two hundred and ten bands were collected in the 0.4 to 2.4  $\mu\text{m}$  region of the visible and infrared spectrum. Some water absorption channels are discarded, resulting in 191 channels. There are seven information classes, roofs, roads, trails, grass, trees, water, and shadows, in the Washington, DC dataset. The test images of WDC and IPS are depicted in Figure 3(a) and 3(b), respectively. Each of them is a portion of the original image.



(a) Washington DC Mall

(b) Indian Pines Site

Figure 3. The test images of a portion of the Washington DC Mall and Indian Pines Site data are displayed in (a) and (b), respectively.

## 4.2 Experimental Design

To evaluate the performance of the proposed algorithms, a portion of the original WDC image and IPS image are selected as a test field, as shown in Figure 3. Three different cases, each class with 20 (case I:  $N_i = 20 < N < d$ ), 40 (case II:  $N_i = 40 < d < N$ ) and 300 (case III:  $d < N_i = 300 < N$ ) training samples are investigated to discover the effect of training sample size in the experiments. In all cases, the test sample size of each class is 100. Of the three cases, the ill-posed (cases I) and poorly posed (cases II) classification problems [29] are included, which are challenging cases in the field of pattern recognition. At each experiment, ten training and testing data sets are randomly selected for computing the average of testing data accuracies of different algorithms.

Two other methods, NWFE and LDA, are used to compare the effectiveness of the proposed NFFE. In this study, the 1NN is implemented in PRTools [30] and the soft-margin SVM classifier is in LIBSVM [31]. For the soft-margin SVM classifier, a parameter  $C$  is used to control the trade-off between the margin and the size of the slack variables. The kernel function employed in SVM is RBF kernel function with a parameter  $\sigma$ . We use the five-fold cross validation to find the best  $C$  and  $\sigma$  within the given sets  $\{2^{-5}, 2^{-3}, \dots, 2^{15}\}$  and  $\{2^{-15}, 2^{-13}, \dots, 2^3\}$ , respectively (suggested by [32]). The values of  $k_1$  and  $k_2$  for estimating the membership grades and the weighted local mean in NFFE are selected by some trial and error in this study. For the WDC data set, the values of  $k_1$  and  $k_2$  are fixed at 5. For the IPS dataset, they are fixed at 3.

## 4.3 Experimental Results

The highest averaged accuracies and their corresponding standard deviations of the feature extraction methods in the experiments are listed in Table I. We have the following findings.

- 1) From an overall viewpoint, NFFE and NWFE make significant improvements on the two data sets, particularly in cases I and II.
- 2) In most of the cases, the number of features required for achieving better classification performance is more than  $L - 1$  for NFFE and NWFE on the two data sets.
- 3) SVM and 1NN classifiers with NFFE features achieve better performances than with other algorithms' features, particularly in cases I and II. In other words, NFFE is more robust than other methods on various training sample sizes.
- 4) LDA obviously yields ill performances in case I on both data sets because it suffers from the singularity problem NFFE and NWFE do not

have the problem due to they include the regularization technique.

- 5) LDA performs poorly on the IPS data set even in case III. By analyzing the mentioned deficiencies of LDA, this phenomenon shows that the IPS data set is possibly not normally distributed. However, NFFE and NWFE still achieve sound results. No matter which classifier we employed, the accuracy difference between LDA and NFFE reach 9.5%.

Table I. The highest average classification accuracies (in %) and their corresponding standard deviations (in %) achieved by using 1NN and SVM classifiers with features extracted by LDA, NWFE and NFFE. The number in the parenthesis represents the number of features achieving the highest accuracies in the experiments.

Data Set	FE	Classifier	$N_i = 20$	$N_i = 40$	$N_i = 300$
			acc±std ( $p$ )	acc±std ( $p$ )	acc±std ( $p$ )
WDC	None	1NN	84.3±1.6	87.5±1.4	94.2±0.8
	NFFE		<b>90.6±1.5</b> (11)	<b>93±1.0</b> (13)	<b>96.7±0.5</b> (13)
	NWFE		88.8±1.8 (4)	91.2±1.2 (7)	95.1±0.6 (8)
	LDA		67.3±2.6 (6)	88.2±1.9 (6)	94.9±0.7 (6)
	None	SVM	84.7±1.3	87.6±0.6	95.1±0.5
	NFFE		<b>91.1±0.9</b> (8)	<b>93±1.0</b> (10)	<b>96.1±0.4</b> (14)
	NWFE		89.3±1.2 (4)	92±0.6 (4)	95.7±0.5 (5)
	LDA		66.9±2.6 (6)	88.1±1.7 (6)	94.7±0.4 (6)
IPS	None	1NN	66.7±1.0	71.9±1.0	82.3±1.0
	NFFE		<b>80.6±1.5</b> (8)	<b>85.2±1.4</b> (11)	<b>93.5±0.6</b> (12)
	NWFE		79.7±2.9 (8)	82.7±1.9 (8)	92.8±0.9 (8)
	LDA		62.2±2.2 (7)	65.3±1.4 (7)	84±0.9 (7)
	None	SVM	75.6±0.9	80±1.7	89.7±0.4
	NFFE		<b>81.4±1.9</b> (8)	<b>85.1±1.5</b> (14)	<b>92.1±0.4</b> (15)
	NWFE		79.4±2.2 (15)	82.8±1.7 (14)	90.4±0.9 (12)
	LDA		61.1±1.5 (7)	64.5±1.4 (7)	82.6±1.1 (7)

Since the feature adjustment procedure is not included in LDA, NWFE and NDA, its influences on NFFE over WDC and IPS data sets by using 1NN and SVM classifiers are investigated and demonstrated in Figure 4. The results with and without feature adjustment procedure in cases I and II are demonstrated. The red lines in these plots denote that the feature adjustment procedure is not adopted. Some remarks can be made as follows:

- 1) Undoubtedly, the feature adjustment procedure is definitely important for obtaining better re-

sults.

- 2) In DC+1NN cases, the accuracies of NFFE\_NonAdj have the tendency to decline with the increase of dimensionality, but it does not occur when feature adjustment is adopted.
- 3) Evident peaks exist when the dimensionality is near ten for NFFE as demonstrated in Figure 4(c), Figure 4(d), and Figure 4(g).
- 4) The Figure 4(f) shows that a higher accuracy can be obtained by using SVM classifier with more than fifteen features. The necessity of specifying more than  $L - 1$  features is verified again.

Some classified images of DC and IPS by using 1NN and SVM classifiers with features extracted by different feature extraction algorithms in cases II ( $N_i = 40$ ) and III ( $N_i = 300$ ) are demonstrated in Figures 5 and 6, respectively. The DC thematic maps are obtained by using 1NN classifier, and the IPS thematic maps are obtained by using SVM classifier. Here  $p$  denotes the number of features achieving the highest accuracies in the experiments, as shown in Table I. On DC image classification, NFFE evidently obtains the best visual effect than the other methods. As shown in Figure 5, NFFE has excellent classification in “grass” and “tree” parts. From Figure 6, we can find that NFFE+SVM achieves the best visual effect. Clearly, in Figure 6(e), all parts classified by NFFE+SVE are better than by NFFE+SVM, LDA+SVM and single SVM, particularly in the areas of “Corn-notill”, “Soybeans-notill” and “Soybeans-min”. A large portion of the “Soybeans-min” part at the top-right corner is misclassified into “Soybeans-clean” by the single SVM, but this area is classified well by NFFE+SVM and NFFE+SVE. LDA+SVM misclassifies a large portion of the “Soybeans-min” part in the center of the map to “Corn-notill”, but NFFE+SVM and NFFE+SVE misclassify only a small portion of that part to “Soybeans-notill”.

## 5. Conclusions and Future Work

The objective of the study is to demonstrate the construction of a powerful nonparametric feature extraction algorithm. We demonstrate that a powerful nonparametric feature extraction algorithm consists of three constituents, including the construction of sound scatter matrices, the employment of a suitable regularization, and the adjustment of the extracted features. In addition, we recommend adopting the feature adjustment procedure because it is of theoretical and practical importance for the design of such a method. This procedure is of practical importance for reaching better results.

NFFE is theoretically suitable for handling linear problems but for the nonlinear problems. The kernel method [26], [33] provides a framework for constructing

the nonlinear version of NFFE to extract nonlinear features. Future research could be conducted on the construction and investigation of the kernel version of NFFE. Moreover, in recent years the type-2 fuzzy set theory [34] has become an important theme in fuzzy systems. The type-2 fuzzy set theory is more capable of handling uncertainty of problem. In [35], type-2 fuzzy classifiers were applied to land cover classification problem. It might be beneficial to introduce the type-2 fuzzy set theory to our work as well.

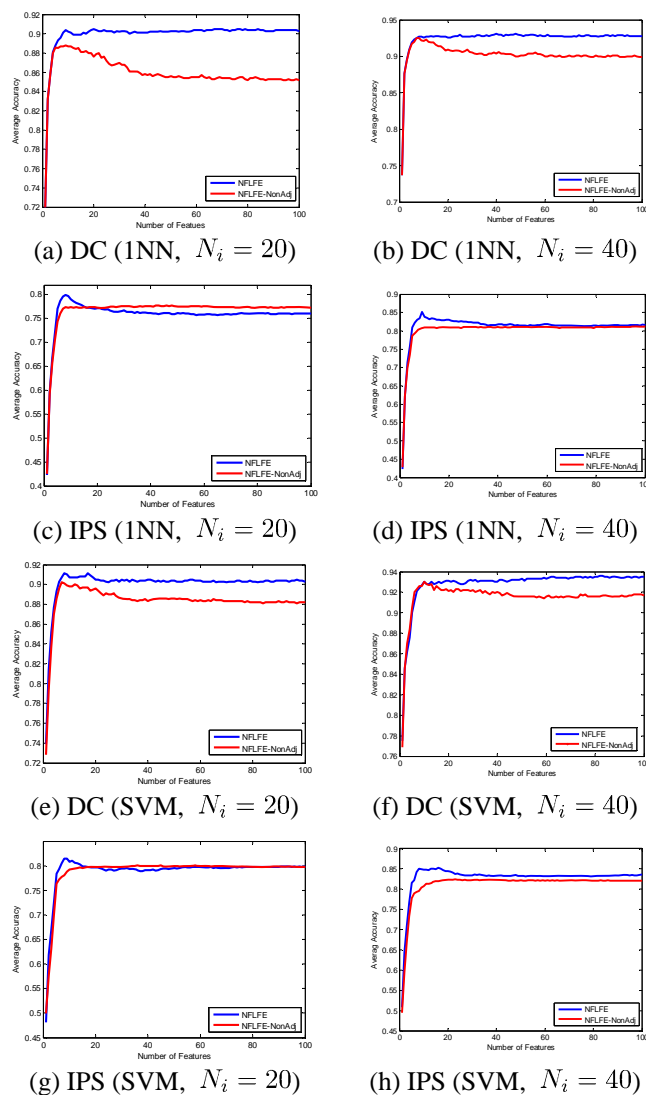


Figure 4. The influence of the feature adjustment procedure on the performances of the proposed NFFE in cases I and II on the DC and IPS data sets.

## Acknowledgment

The authors would like to thank Prof. Landgrebe for providing the Indian Pines and Washington DC Mall data sets.

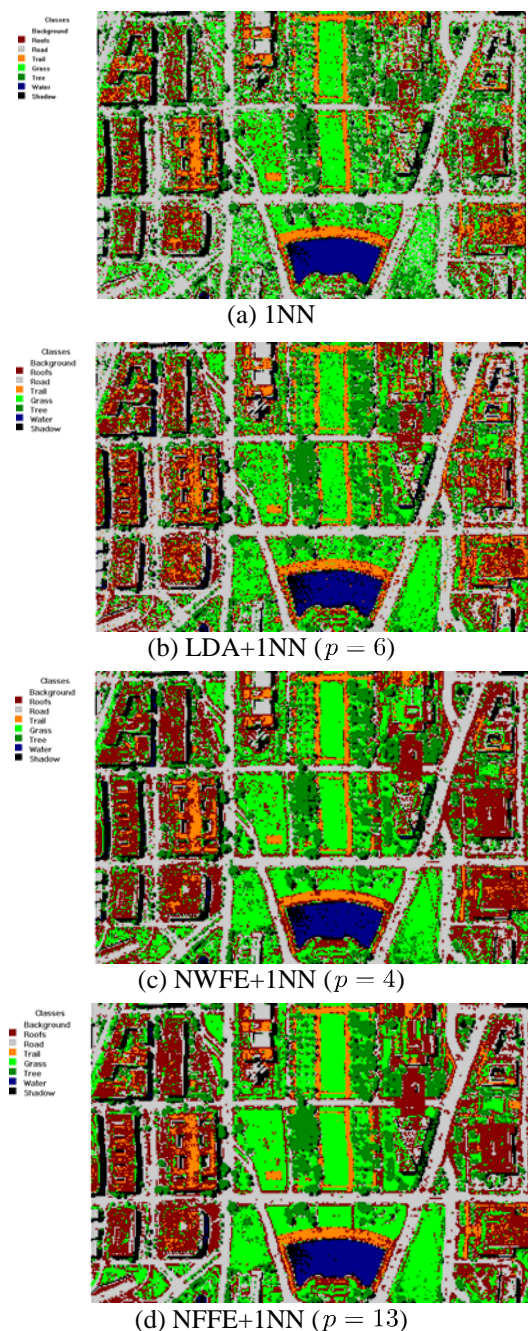


Figure 5. Thematic maps resulting from the classification of the area of Figure 3(a) in case II ( $N_i = 40$ ). (a) to (d) are the results of 1NN, LDA+1NN, NFFE+1NN and NFFE+1NN, respectively.

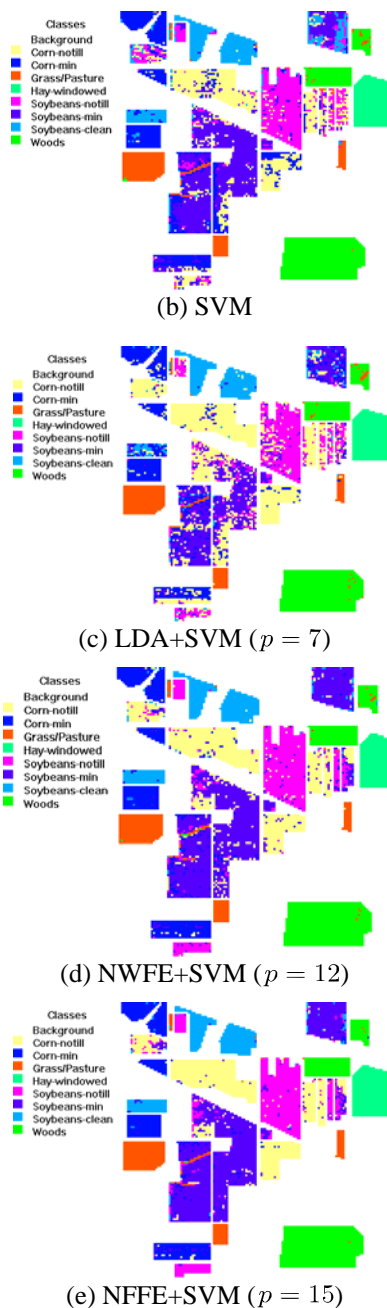
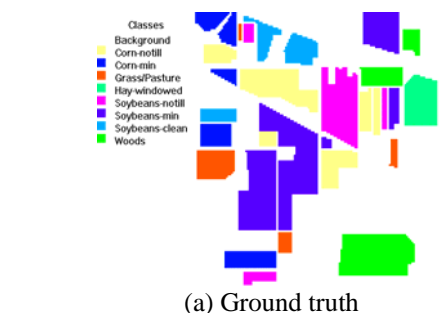


Figure 6. Thematic maps resulting from the classification of the area of Figure 3(b) in case III ( $N_i = 300$ ). (a) is the ground truth of the area with eight classes, and (b) to (e) are the results applying single SVM, LDA+SVM, NFFE+SVM, NFFE+SVM, respectively.



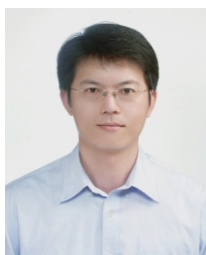
### References

- [1] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, NJ: Wiley, 2003.
- [2] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55-63, 1968.
- [3] P. K. Varshney and M. K Arora. *Advanced image processing techniques for remotely sensed hyperspectral data*, Springer, New York, 2004.

- [4] R. O. Duda, P. E. Hart, and D. G. Stock, *Pattern Classification*, John Wiley & Sons, 2nd, 2001.
- [5] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, San Diego, CA:Academic Press, 1990.
- [6] F. van der Heiden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*, Chichester: John Wiley & Sons, 2004.
- [7] B. C. Kuo and D. A. Landgrebe, "Nonparametric weighted feature extraction for classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 5, pp. 1096-1105, May 2004.
- [8] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252-264, 1991.
- [9] K. Fukunaga and J. M. Mantock, "Nonparametric discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, pp. 671-678, 1983.
- [10] B. C. Kuo and K. Y. Chang, "Feature extractions for small sample size classification problem", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 3, pp. 756-764, March 2007.
- [11] P. 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] J. Wang and C. I. Chang, "Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 6, pp.1586-1600, June 2006.
- [14] H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, pp.165-175, 1989.
- [15] P. F. Hsieh, D. S. Wang, and C. W. Hsu, "A linear feature extraction for multiclass classification problems based on class mean and covariance discriminant information, extraction," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 223-235, Feb. 2006.
- [16] X. Song, G. Fan, and M. Rao, "Automatic CRP mapping using nonparametric machine learning approaches," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 888-897, 2005.
- [17] J. A. Richards, "Analysis of remotely sensed data: the formative decades and the future," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 422-432, 2005.
- [18] D. Landgrebe, "Multispectral land sensing: Where from, where to?" *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 414-421, 2005.
- [19] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 480-491, 2005.
- [20] M. M. Dundar and D. Landgrebe, "Toward an optimal supervised classifier for the analysis of hyperspectral Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 1, pp. 271-277, 2004.
- [21] B. C. Kuo and D. A. Landgrebe, "A covariance estimator for small sample size classification problems and its application to feature extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 4, pp. 814-819, 2002.
- [22] J. P. Hoffbeck and D. A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Trans. on Pattern Analysis & Machine Intelligence*, vol. 18, no. 7, pp. 763-767, 1996.
- [23] S. Tadjudin and D. A. Landgrebe, "Covariance estimation with limited training samples," *IEEE Transaction on Geoscience and Remote Sensing*, vol. 37, no. 4, pp. 2113-2118, July 1999.
- [24] J. M. Keller, M. R. Gray, and J. A. Givens, Jr., "A fuzzy k-nearest neighbor algorithm", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 15, no. 4, pp. 580-58, July/August, 1985.
- [25] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer-Verlag, 2001.
- [26] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.
- [27] R. O. Green, et al., "Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS)," *Remote Sensing of Environment*, vol. 65, pp.227-248.
- [28] R. G. Resmini, et al., "Mineral mapping with hyperspectral digital imagery collection experiment (HYDICE) sensor data at cuprite," Nevada, U. S. A. *International Journal of Remote Sensing*, vol. 18, no. 7, pp. 1553-1570, 1997.
- [29] 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.
- [30] R. P. W. Duin, P. Juszczak, P. Paclik, E. Pekalska,

D.de Ridder, D.M.J. Tax, S. Verzakov, *PRTTools4.1, A Matlab Toolbox for Pattern Recognition*, Delft University of Technology, 2007.

- [31] C. C. Chang and C. J. Lin, "LIBSVM : a library for support vector machines," [Online]. Available: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [32] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," [Online]. Available: <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>
- [33] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, New York, 2004.
- [34] J. M. Mendel, *Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions*, Prentice-Hall, Upper-Saddle River, NJ, 2001.
- [35] L. A. Lucas, T. M. Centeno, and M. R. Delgado, "Land Cover Classification Based on General Type-2 Fuzzy Classifiers," *International Journal of Fuzzy Systems*, vol. 10, no. 3, pp.207-216, Sep. 2008.



**Jinn-Min Yang** received the B.S. and M.S. degree from the National Taichung Teachers College, Taichung, Taiwan, R.O.C., in 1994 and 2000, respectively, and the Ph.D. degree from the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan in 2010. He is with the Department of Mathematics

Education at National Taichung University of Education, Taichung, Taiwan. His research interests include fuzzy systems, pattern recognition, remote sensing, machine learning, and e-Learning.



**Pao-Ta Yu** received the B.S. degree in mathematics from National Taiwan Normal University in 1979, the M.S. degree in computer science from National Taiwan University, Taipei, Taiwan, in 1985, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, in 1989. Since 1990, he has been with the Department of Computer Science and Information Engineering at National Chung Cheng University, Chiayi, Taiwan, R.O.C., where he is currently a Professor. His research interests include e-Learning, neural networks and fuzzy systems, nonlinear filter design, intelligent networks, and XML technology.



**Bor-Chen Kuo** received the B.S. and M.S. degrees from National Taichung Teachers College, Taichung, Taiwan, R.O.C., in 1993 and 1996, respectively, and the Ph.D. degree from school of electrical and computer engineering, Purdue University, West Lafayette, IN, in 2001. He is currently a Professor in the Graduate Institute of Educational Measurement and Statistics at the National Taichung University of Education, Taiwan. His research interests are pattern recognition, remote sensing, image processing, and nonparametric functional estimation.



**Ming-Hsiang Su** received the B.S. degree in computer science from the Tunghai University, Taichung, Taiwan, in 2001, the M.S. degree in management information systems from the National Pingtung University of Science and Technology, Pingtung, Taiwan, in 2003. He is currently a doctoral candidate in the Department of Computer Science

and Information Engineering, National Chung Cheng University, Chiayi, Taiwan. His research interests include e-learning, artificial intelligence and machine learning.