

A Semisupervised Feature Extraction Method Based on Fuzzy-type Linear Discriminant Analysis

Hui-Shan Chu

Graduate Institute of Educational Measurement and
Statistics, National Taichung University of Education,
Taichung, Taiwan, R.O.C.
Roxanne90@gmail.com

Bor-Chen Kuo

Graduate Institute of Educational Measurement and
Statistics, National Taichung University of Education,
Taichung, Taiwan, R.O.C.
kbc@mail.ntcu.edu.tw

Cheng-Hsuan Li

Institute of Electrical Control Engineering,
National Chiao Tung University,
Hsinchu, Taiwan, R.O.C.
ChengHsuanLi@gmail.com

Chin-Teng Lin

Institute of Electrical Control Engineering,
National Chiao Tung University,
Hsinchu, Taiwan, R.O.C.
ctlm@mail.nctu.edu.tw

Abstract—Linear discriminant analysis (LDA) is a commonly used feature extraction (FE) method to resolve the Hughes phenomenon for classification. The Hughes phenomenon (also called the curse of dimensionality) is often encountered in classification when the dimensionality of the space grows and the size of the training set is fixed, especially in the small sampling size problem. Recent studies show that the spatial information can greatly improve the classification performance. Hence, for hyperspectral image classification, it is not only necessary to use the available spectral information but also to exploit the spatial information. In this paper, a semisupervised feature extraction method which is based on the scatter matrices of the fuzzy-type LDA and uses the semi-information is proposed. The experimental results on two hyperspectral images, the Washington DC Mall and the Indian Pine Site, show that the proposed method can yield a better classification performance than LDA in the small sampling size problem.

Keywords—feature extraction; linear discriminate analysis

I. INTRODUCTION

Linear discriminant analysis (LDA) [4] is a commonly used feature extraction (FE) method to resolve the Hughes phenomenon (also called the curse of dimensionality) which means that when the dimension of the space grows and the size of the training set is fixed, the classification accuracy reach a maximum for a given size, then it decreases for the hyperspectral image classification. Recently, a fuzzy-type LDA (FLDA), an unsupervised version of linear discriminant analysis (UFLDA), based on the membership values is proposed [3]. The membership values of the scatter matrices in FLDA are obtained from the cluster result of fuzzy linear discriminant clustering (FLDC). Then the generalized eigenvectors form the transformation matrix to transform the samples in the space with lower dimensionality.

Moreover, the spatial information by considering semi-labels can greatly improve the accuracy of the hyperspectral image classification [1]-[2]. Using the semi-labels of the

unknown samples can not only increase the training samples but also decrease the imprecise estimation of samples, which have very similar spectral properties, and have difficulty distinguishing the unlabeled patterns [1]-[2].

Therefore, in this study, we want to integrate the scatter matrices of FLDA and semi-information to develop spatial scatter matrices. The membership values of the scatter matrices of FLDA are determined by the posteriors of every pixel in the image from the previous classification results. After obtaining the within- and between-cluster scatter matrices based on the spatial information, combine with the scatter matrices of LDA to form semi-scatter matrices. Therefore, in this paper, a LDA based on semi-information (SLDA) is proposed. The experimental results on the Washington DC Mall image and the Indian Pine Site image data sets show that the proposed method can yield a better classification performance than LDA when sampling size is small.

The study is organized as followed. The LDA and its unsupervised version will be introduced in section II. Section III describes our proposed method, SLDA, and the experimental results are shown in Section IV. Finally, some conclusions are addressed in Section V.

II. LINEAR DISCRIMINANT ANALYSIS AND ITS UNSUPERISED VERSION

A. Linear Discriminant Analysis (LDA)

LDA [4] uses the mean vector and covariance matrix of each class, to formulate within-class, between-class, and mixture-class scatter matrices. Based on the Fisher criterion, the LDA method finds features such that the ratio of the between-class scatter to the average within-class scatter is maximized, in a lower dimensional space. By applying the concept of class scattering to class separation, the Fisher criterion, takes the large values from samples when they are well clustered around their mean within each class, and the

clusters of the different classes are well separated. The formula is stated as the following:

Suppose $H_i = \{x_1^{(i)}, \dots, x_{N_i}^{(i)}\} \subset R^d$ were the set of samples in class i , N_i were the number of samples in class i , $i = 1, \dots, L$, and $N = N_1 + \dots + N_L$ were the number of all training samples. The between-class scatter matrix S_b^{LDA} and the within-class scatter matrix S_w^{LDA} would be defined as

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

and

$$S_w^{LDA} = \sum_{i=1}^L \sum_{j=1}^{N_i} \frac{1}{N} (x_j^{(i)} - m_i)(x_j^{(i)} - m_i)^T \quad (2)$$

where m_i were the class mean defined by $m_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_j^{(i)}$ and

$$m = \frac{1}{N} \sum_{i=1}^L \sum_{j=1}^{N_i} x_j^{(i)}$$
 representing the total mean.

The optimal features are determined by optimizing the Fisher criterion given by $J_{LDA} = \text{tr}[(S_w^{LDA})^{-1} S_b^{LDA}]$. This is equivalent to solving the generalized eigenvalue problem, $S_b^{LDA} v_z = \lambda_z S_w^{LDA} v_z$, $z = 1, \dots, d$ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ where the extracted eigenvectors are used to form the transformation matrix of LDA, i.e., the transformation matrix from the original space to the reduced subspace is defined by: $A^{LDA} = [v_1, v_2, \dots, v_p]$. The Fisher criterion J_{LDA} is able to detect the capacity of the separability for the transformed training samples.

B. Fuzzy-type Linear Discriminant Analysis (FLDA)

In the subsection, an unsupervised LDA, fuzzy-type LDA (FLDA), using the concept of membership values is described [3]. Let X be a hyperspectral d -dimensional image of size $N \times z$ pixels, and $\{x_1, \dots, x_{N \times z}\} \subset R^d$ be all samples in the image. The between scatter matrix S_b^{FLDA} and the within scatter matrix S_w^{FLDA} as follows:

$$S_b^{FLDA} = \sum_{i=1}^L \frac{\sum_{t=1}^{N \times z} u_{it}}{N \times z} (m_i - m)(m_i - m)^T \quad (3)$$

and

$$S_w^{FLDA} = \sum_{i=1}^L \sum_{t=1}^{N \times z} \frac{u_{it}}{N \times z} (x_t - m_i)(x_t - m_i)^T \quad (4)$$

where u_{it} means the membership value of the sample x_t in class i ,

$$m_i = \sum_{t=1}^{N \times z} \frac{u_{it}}{\sum_{k=1}^L u_{ik}} x_t$$

is the class mean, and $m = \frac{1}{N \times z} \sum_{t=1}^{N \times z} x_t$ represents the total mean.

The weight $\sum_{t=1}^{N \times z} u_{it} / (N \times z)$ indicates the influence of the samples in class i . If clustering results show that the number of samples in cluster i is larger than in cluster k , then

$$\sum_{t=1}^{N \times z} u_{it} / (N \times z) > \sum_{t=1}^{N \times z} u_{kt} / (N \times z),$$

i.e., the influence of m_i is larger than m_k in measuring the between separability. The weight $u_{it} / (N \times z)$ indicates the influence of the t -th sample in class i . If $u_{it} / (N \times z) > u_{kt} / (N \times z)$, then the k -th sample favors in class i . Hence, the corresponding weights in the within measure is larger.

Furthermore, it is easy to observe that the scatter matrices of LDA are special cases of the scatter matrices of FLDA [3], i.e., $u_{ii} = 1$ and $u_{ki} = 0$, $k \neq i$ when x_t is the training sample in class i . In [3], the clustering performance of FLDA outperforms the clustering performance of principal component analysis (PCA) and independent component analysis (ICA).

III. A SPECTRAL AND SPATIAL FEATURE EXTRACTION METHOD

In this section, we propose a semi-information based LDA (SLDA) algorithm. The purpose of incorporating the semi-information in feature extraction is to increase the useable samples. In this study, the posteriori probabilities by applying the Gaussian classifier are used to form the weights in the scatter matrices of FLDA. Then the semi-information then can be both used to adjust the extracted direction, and the classification performance will be improved.

After applying the Gaussian classifier [4], the class-conditional probability density function of the unknown sample x_j is

$$p(x_j | i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2} (x_j - m_i)^T \Sigma_i^{-1} (x_j - m_i)\right)$$

where $m_i = \frac{1}{N_i} \sum_{t=1}^{N_i} x_t^{(i)}$ is the mean of class i ,

$$\Sigma_i = \frac{1}{N_i} \sum_{t=1}^{N_i} (x_t^{(i)} - m_i)(x_t^{(i)} - m_i)^T$$

is the $d \times d$ covariance matrix, and $|\Sigma_i|$ denotes the determinant of Σ_i . Then x_j is distinguished to class i , if

$$p(x_j | i)p(i) \geq p(x_j | k)p(k)$$

for all $k \neq i$, where $p(i)$ is the priori probability of class i .

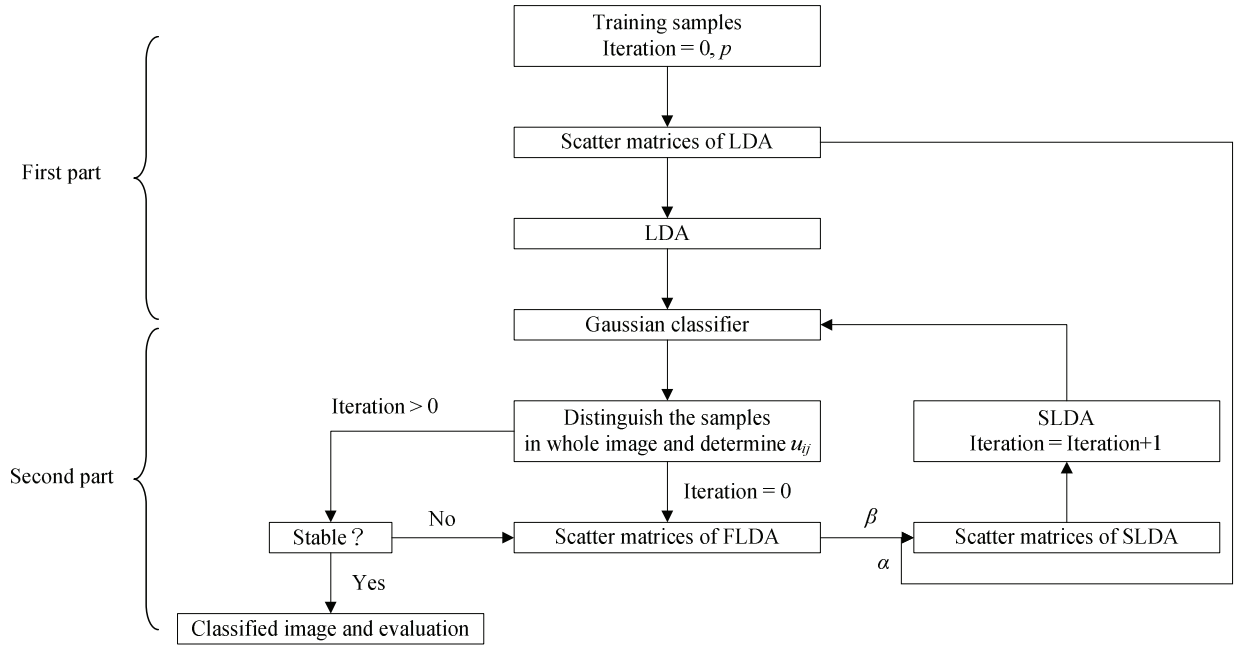


Figure 1. The framework of the proposed method, SLDA, which is based on semi-information.

In SLDA, the membership values of the semi-scatter matrices are defined as the posterior probabilities. According to the Bayes rule, the membership value u_{iq} of the unknown sample x_q , is defined as

$$u_{iq} = p(i | x_q) = \frac{p(x_q | i)p(i)}{\sum_{i=1}^L p(x_q | i)p(i)}.$$

for $i = 1, \dots, L$, and $q = 1, \dots, N \times z$.

The between scatter matrix S_b^{SLDA} and the within scatter matrix S_w^{SLDA} are defined as follows:

$$S_b^{SLDA} = \alpha S_b^{LDA} + \beta S_b^{FLDA} \quad (5)$$

and

$$S_w^{SLDA} = \alpha S_w^{LDA} + \beta S_w^{FLDA} \quad (6)$$

where α and β are two parameters to control the influences of the scatter matrices of LDA and semi-scatter matrices of FLDA. In this paper, the k -fold cross validation method is used to determined the proper parameters, α and β .

The features of SLDA can be obtained by solving the following generalized eigenvalue problem: $S_b^{SLDA} \mathbf{v}_z = \lambda_z S_w^{SLDA} \mathbf{v}_z$. The extracted eigenvectors form the transformation matrix of SLDA, i.e.,

$$A^{SLDA} = [v_1, v_2, \dots, v_p].$$

One can find that if $\beta = 0$, then

$$A^{SLDA} = A^{LDA}.$$

TABLE I. THE ALGORITHM OF SLDA

Step 1:	Set the number of extracted features $p \leq d$. Use the transformation A^{LDA} to project the samples in the image and obtain the semi-labels by the Gaussian classifier.
Step 2:	Compute u_{it} for every sample x_t in the whole image.
Step 3:	Compute S_b^{FLDA} and S_w^{FLDA} .
Step 4:	Compute S_b^{SLDA} and S_w^{SLDA} by multiplying the parameters α and β .
Step 5:	Solve the eignevalue problem $S_b^{SLDA} \mathbf{v}_z = \lambda_z S_w^{SLDA} \mathbf{v}_z$, $z = 1, \dots, d$ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$.
Step 6:	The transformation function from the original space to the reduced subspace is defined by $A^{SLDA} = [v_1, v_2, \dots, v_p]$.
Step 7:	Use the transformation A^{SLDA} to project the samples in the image and obtain the new semi-labels by the Gaussian classifier.
Step 8:	Repeat steps 2 and 7 until the iteration and the above iteration is below a certain tolerance value or it's no longer decreasing.

Fig. 1 is the framework of the proposed method, SLDA, which is based on semi-information. Given the training samples and determine the reduced dimensionality p . In the first part, the scatter matrices of LDA are computed and LDA is used for feature extraction. After applying the Gaussian classifier, the semi-labels and membership values can be obtained, and the scatter matrices of FLDA in the second part can be computed. By multiplying the parameters α and β , the scatter matrices of SLDA can be computed. After applying the SLDA and the Gaussian classifier, the new semi-labels and membership values can be obtained. If the change rate of semi-

labels is less than a threshold, then stop the iteration. Otherwise, continue the second part until the change rate of semi-labels is less than a threshold. The algorithm of SLDA is described in the Table I.

IV. EXPERIMENTAL RESULTS

A. Experimental Data and Design

In this study, two real data sets are applied to evaluate the classification performances. They are the Indian Pine site (IPS): a mixed forest/agricultural site in Indiana [5] and the Washington, D.C. Mall hyperspectral image [5] as an urban site.

The IPS dataset was gathered by a sensor known as the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). This dataset was obtained from an aircraft flown at 65000ft. altitude and operated by the NASA/Jet Propulsion Laboratory, with a size of 145×145 pixels and has 220 spectral bands measuring approximately 20m across the ground. The grayscale IR image and ground truth of IPS are shown in Figs. 2 and 3, respectively. There are 9 different land-cover classes available in the original ground-truth. 9 categories, Corn – no till, Corn – min till, Grass / Pasture, Grass / Trees, Hay – windrowed, Soybean – no till, Soybean – min till, Soybean – clean till, Woods, would be used in our experiments.

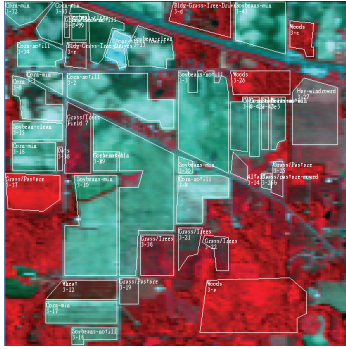


Figure 2. A portion of the Indian pine site image with a size of 145×145 pixels.



Figure 3. The ground truth of Indian pine site dataset.

The second data set, the Washington, D.C. Mall (DC) from an urban area, is a Hyperspectral Digital Imagery Collection Experiment (HYDICE) airborne hyperspectral data flightline over the Washington, DC Mall. Two hundred and ten bands were collected in the $0.4\text{--}2.4\ \mu\text{m}$ region of the visible and infrared spectrum. Some water absorption channels are

discarded, resulting in 191 channels. The dataset is available in the student CD-ROM of (Landgrebe, 2003). In our experiment, there are 7 classes, grass (class 1), tree (class 2), roof (class 3), water (class 4), road (class 5), trail (class 6), and shadow (class 7). The grayscale IR image of a portion of the image and the corresponding seven categories are shown in Fig. 4.

No.	Category
1	grass
2	tree
3	roof
4	water
5	road
6	trail
7	shadow

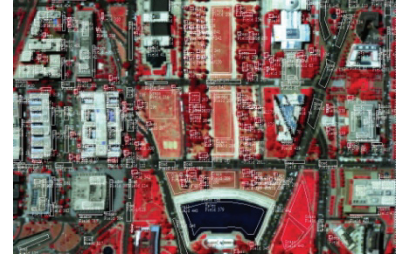


Figure 4. The false-color IR image of a portion of Washington, D.C. Mall image with a size of 205×307 pixels. There are seven categories, grass, tree, roof, water, road, trail, and shadow.

In the experiment, for investigating the influence of training sample size to the dimensionality, three distinct subsets, which N_i is the number of each class training samples, d is the number of dimension, and N is the number of total training samples. $N_u=20 < N < d$ (case 1), $N_u=40 < d < N$ (case 2), and $d < N_u=300 < N$ (case 3), will be discussed in these two hyperspectral image data sets. For example, the Indian Pine site (IPS) dataset. In case 1, $N_u=20 < N=180 < d$ is a so called ill-posed classification situation, which means data dimensionality exceeds the total number of independent training samples in every class. In case 2, $N_u=40 < d < N$ is a so called poorly-posed classification situation, which means data dimensionality is greater than or comparable to the number of (independent) per-class representative training samples, but smaller than the total number of representative samples. In case 3, there are enough independent training samples. The MultiSpec [5] was used to select training and testing samples (100 testing samples per class) randomly in our experiments [6]–[8].

In the experiments, the classification performances by applying LDA and SLDA with the Gaussian classifier are compared. The grid search with k -fold cross validation is used to find the proper parameters α and β within a set $\{0, 0.1, 0.2, \dots, 2\}$.

B. Results

In the experimental results, the classification accuracies in percentages for two different data sets are showed in Table II–IV. Because of LDA only can extract $L-1$ (number of classes minus one) features. For example, the Indian Pine site (IPS) dataset has 9 categories, so LDA only can extract 8 features. As result of experiment we can know that the number of dimensionality with the classification accuracies has weak correlation. The classification performances by applying our proposed SLDA are higher than the classification performances by applying LDA in cases 1 and 2. However, in case 3 (Table IV), the accuracies by applying our proposed method are similar to that by applying LDA. Hence, in the small sample

size problem, the proposed method can yield a better classification performance than LDA.

Fig. 5-7 and 8-10 show the classification maps of the IPS and DC data sets by applying LDA and SLDA with the Gaussian classifier in three cases, respectively. From Fig. 5-6 and 8-9, i.e., case 1 and case 2, the proposed method SLDA can not only increase the classification performance but also decrease the imprecise estimation of samples, especially, “Woods,” “Grass/trees,” “Grass/pasture,” “Soybeans-clean till,” and “Hay-windowed” for IPS dataset, and all classes for DC dataset. Similarly, the classification maps are similar by applying LDA and SLDA for these two data sets except the “trail” in DC dataset. For this class (“trail”), SLDA outperforms LDA.

TABLE II. THE ACCURACY IN PERCENTAGES FOR TWO DIFFERENT DATA SETS IN CASE 1, $N_r=20$. THE NUMBER IN BRACKET IS THE REDUCED DIMENSIONALITY

FE	IPS	DC
LDA	55.89% (7)	47.71% (3)
SLDA	69.56% (8)	90.86% (4)

TABLE III. THE ACCURACY IN PERCENTAGES FOR TWO DIFFERENT DATA SETS IN CASE 2, $N_r=40$. THE NUMBER IN BRACKET IS THE REDUCED DIMENSIONALITY

FE	IPS	DC
LDA	69.00% (8)	86.29% (6)
SLDA	78.89% (8)	92.00% (4)

TABLE IV. THE ACCURACY IN PERCENTAGES FOR TWO DIFFERENT DATA SETS IN CASE 3, $N_r=300$. THE NUMBER IN BRACKET IS THE REDUCED DIMENSIONALITY

FE	IPS	DC
LDA	87.78% (8)	96.14% (5)
SLDA	87.56% (8)	94.71% (6)

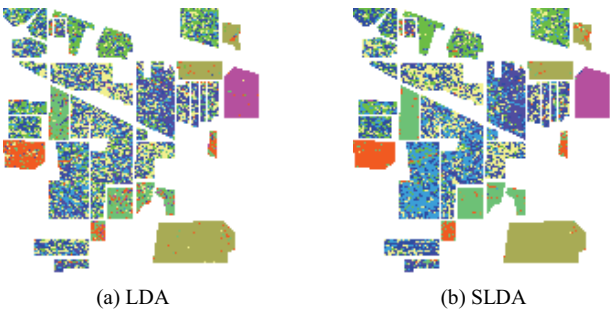


Figure 5. (a) and (b) are the classification maps of the IPS dataset by applying LDA and SLDA with the Gaussian classifier in case 1 ($N_r=20$), respectively.

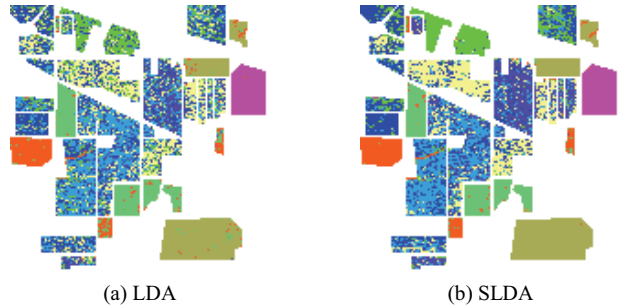


Figure 6. (a) and (b) are the classification maps of the IPS dataset by applying LDA and SLDA with the Gaussian classifier in case 2 ($N_r=40$), respectively.

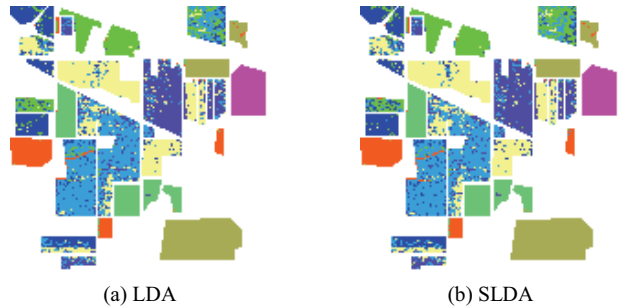


Figure 7. (a) and (b) are the classification maps of the IPS dataset by applying LDA and SLDA with the Gaussian classifier in case 3 ($N_r=300$), respectively.

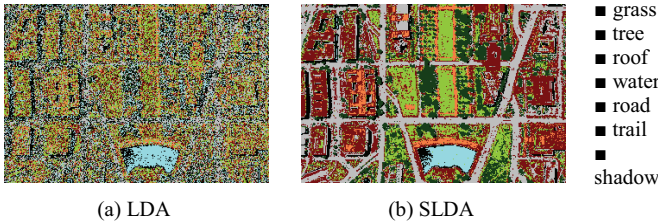


Figure 8. (a) and (b) are the classification maps of the DC dataset by applying LDA and SLDA with the Gaussian classifier in case 1 ($N_r=20$), respectively.

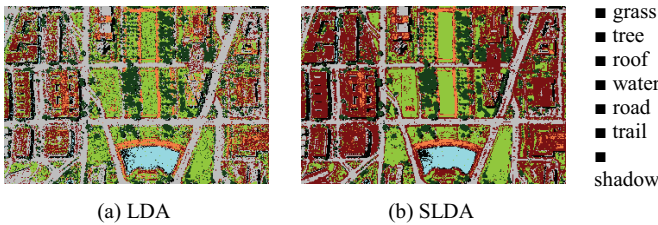


Figure 9. (a) and (b) are the classification maps of the DC dataset by applying LDA and SLDA with the Gaussian classifier in case 2 ($N_r=40$), respectively.

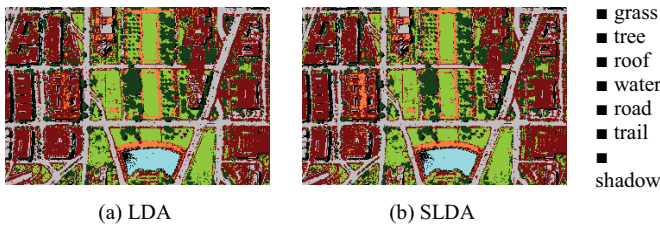


Figure 10. (a) and (b) are the classification maps of the DC dataset by applying LDA and SLDA with the Gaussian classifier in case 3 ($N_u=300$), respectively.

V. CONCLUSION

In this paper, a semi-information based LDA (SLDA) algorithm is proposed. The posteriors after applying the Gaussian classifier were used to determine the membership values of FLDA. Combining the scatter matrices of LDA and semi-scatter matrices of FLDA, the scatter matrices of SLDA are computed. According to the experiments on the Washington DC Mall image and the Indian Pine Site image data sets, SLDA can yield a better classification performance than LDA when sampling size is small.

ACKNOWLEDGMENT

The authors would like to thank Dr. Landgrebe for providing the Indian Pines site and The Washington DC Mall data sets. This work was supported in part by the National

Science Council, Taiwan, under Grant NSC 98-2221-E-142-005.

REFERENCES

- [1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed., San Diego, CA: Academic, 1990.
- [2] C.-H. Li, B.-C. Kuo, and C.-T. Lin, "LDA-based Clustering Algorithm and Its Application to an Unsupervised Feature Extraction," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 1, pp.152-163, Feb. 2011.
- [3] Q. Jackson and D.A. Landgrebe, "Adaptive Bayesian Contextual Classification Based on Markov Random Fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 11, pp. 2454-2463, 2002.
- [4] B.-C. Kuo, C.-H. Chuang, C.-S. Huang, and C.-C. Hung, "A nonparametric contextual classification based on Markov random fields," in the *Proceedings First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 2009. WHISPERS '09.
- [5] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*, John Wiley and Sons, Hoboken, NJ: Chichester, 2003.
- [6] B.-C. Kuo, C.-H. Li, and J.-M. Yang, "Kernel Nonparametric Weighted Feature Extraction for Hyperspectral Image Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 4, pp.1139-1155, April 2009.
- [7] J.-M. Yang, P.-T. Yu, and B.-C. Kuo, "A Nonparametric Feature Extraction and Its Application to Nearest Neighbor Classification for Hyperspectral Image Data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp.1279-1293, March 2010.
- [8] J.-M. Yang, B.-C. Kuo, P.-T. Yu, and C.H. Chuang, "A Dynamic Subspace Method for Hyperspectral Image Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 7, pp.2840-2853, July 2010.