Journal of AI and Data Mining
Vol 5, No 1, 2017, 39-53

DOI: 10.22044/jadm.2017.787

Feature extraction of hyperspectral images using boundary semi-labeled samples and hybrid criterion

M. Imani and H. Ghassemian*

Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran.

Received 27 April 2016; Accepted 21 September 2016
**Corresponding author: ghassemi@modares.ac.ir (H. Ghassemian).*

Abstract
Feature extraction is a very important preprocessing step for classification of hyperspectral images. The linear discriminant analysis (LDA) method fails to work in small sample size situations. Moreover, LDA has a poor efficiency for non-Gaussian data. LDA is optimized by a global criterion. Thus, it is not sufficiently flexible to cope with the multi-modal distributed data. In this work, we propose a new feature extraction method, which uses the boundary semi-labeled samples for solving small sample size problems. The proposed method, called the hybrid feature extraction based on boundary semi-labeled samples (HFE-BSL), uses a hybrid criterion that integrates both the local and global criteria for feature extraction. Thus, it is robust and flexible. The experimental results with one synthetic multi-spectral and three real hyperspectral images show the good efficiency of HFE-BSL compared to some popular and state-of-the-art feature extraction methods.

Keywords: *Feature Extraction, Hyperspectral Image, Boundary Samples, Hybrid Criterion, Classification*.

1. Introduction

Hyperspectral imaging has many applications in different fields [1,2]. Analysis of hyperspectral images has a lot of challenges. For instance, due to the impact of the sensor's instantaneous field-of-view and the diversity of the land-cover classes, the presence of mixed pixels is possible. By converting the abundance map into a higher resolution image, the subpixel mapping technique can specify the spatial distribution of different categories at the subpixel scale. In [47], an adaptive subpixel mapping method based on a maximum a posteriori (MAP) model and a winner-take-all class determination strategy has been proposed to improve the accuracy of the subpixel mapping. Classification is one of the most important tasks in a hyperspectral image analysis. The objective of the hyperspectral image classification is to associate each pixel with a proper label [3, 4]. Increasing the spectral bands provided by the hyperspectral imaging technology has brought new potentials and challenges to data analysis. Classification can be done supervised or unsupervised. The unsupervised classification methods are used to solve the site labeling problems without the need for labeled samples. An unsupervised artificial immune classifier has been proposed in [48], which possesses biological properties such as clonal selection, immune network, and immune memory in addition to nonlinear classification properties. Hyperspectral images provide a valuable source of spectral information for class discrimination with lots of details [5]. On the other hand, in order to fully utilize the information contained in the increased features, a large number of training samples is required for the supervised classification of hyperspectral images. Unfortunately, obtaining the training samples is generally expensive and difficult. When the number of training samples is small compared to the number of features, the Hughes phenomenon occurs [6].

Parametric classifiers such as the maximum likelihood (ML) classifier model the probability density functions for individual classes. So, they are often ineffective for the classification of high dimensional data. In order to mitigate the curse of dimensionality and degrade the small sample size problem, there are various ways such as the use of semi-supervised classifiers [8], and non-

parametric classifiers such as kernel-based classifiers [9], for example, support vector machines (SVMs). Another solution to cope with the large number of spectral bands in hyperspectral images is feature reduction. Feature reduction has three main advantages: improvement in classification accuracy, decreasing the computational cost and better understanding of data [11].

In general, feature reduction methods can be categorized into two groups: feature selection and feature extraction [12-14]. A full review of feature reduction methods has been provided in [15]. In general, a feature extraction algorithm provides a feature subset richer than that obtained using a feature selection technique with a higher cost [16]. Some feature selection methods have been represented in [17-19]. In [49], the authors have been proposed a stochastic search strategy inspired by the clonal selection theory in an artificial immune system, where dimensionality reduction is formulated as an optimization problem that searches an optimum with less number of features in a feature space.

Our main focus in this paper is on the feature extraction. The feature extraction methods are divided into two general groups: the supervised ones and the unsupervised ones. There are a variety of methods mentioned for feature extraction in the literature [20-26]. The most popular supervised feature extraction method is linear discriminant analysis (LDA) [27]. Generalized discriminant analysis (GDA) is the nonlinear extension of LDA [28]. LDA requires a much number of training samples for an accurate estimation of the scatter matrices. Thus, it has no good efficiency in small sample size situations. To cope with this problem, there are different approaches. The first approach is the use of unsupervised methods. The unsupervised feature extraction methods require no training samples [29-31]. However, unsupervised feature extraction methods do not consider the class separability, and so, may not be sufficiently appropriate for feature extraction in the classification applications. The second approach for feature extraction, in small sample size situation, is the use of non-parametric feature extraction methods [32-35]. Some feature extraction methods such as decision boundary feature extraction [36] and supervised feature extraction methods based on neural networks [37] need a large training set, and so, they are not efficient methods in small sample size situations. The third approach to cope with the small sample size problem is the semi-supervised approach that uses the ability of both the labeled and unlabeled samples. Semi-supervised discriminant analysis (SDA) has been introduced in [38]. In the SDA method, the labeled samples are used to maximize the discriminating power, while the unlabeled samples are used to maximize the locality preserving power. Some other feature extraction and classification methods to solve the small sample size problem have been proposed in [50]-[53].

LDA is optimized by the global criterion where similarities are measured by distances between the samples and the sample mean vector. Global criterion-based methods such as LDA may perform poorly with multi-modal distributed data. In other words, they only work well with uni-modal distributed samples. In our proposed feature extraction method, the local criterion is added to the global criterion for a discriminant analysis. In the local criterion, similarities are measured by distances between the neighboring pairs. In comparison with the global criterion, the local criterion can treat well with multi-modal distributed data and make it more flexible than the global one. However, the local criterion-based algorithm has a weaker robustness than the global one. The local criterion-based algorithm is more complex than the global one, and hence, more prone to overfitting. In general, the global criterion-based feature extraction method has a stronger robustness and a weaker flexibility, while the local one has a stronger flexibility and a weaker robustness.

In this paper, we use a hybrid criterion that is a combination of the global and local criteria for providing the advantages of both of them. As mentioned earlier, LDA has some difficulties such as the following ones:

- Because of the singularity problem, it does not have good performance in small sample size situations.
- The good performance of LDA is dependent on the Gaussian assumption of data because LDA is optimized by the global criterion, which is robust and non-flexible.

To cope with the difficulties of LDA, we propose a few solutions together. To solve the small sample size problem, we propose to use the high confidence semi-labeled samples. There are three types of samples in an image data. Some samples are labeled-samples, where their labels are obtained by doing field operations or thought of experts. Most samples are unlabeled, and so, it is not specified which class they belong to. The third group is semi-labeled samples.

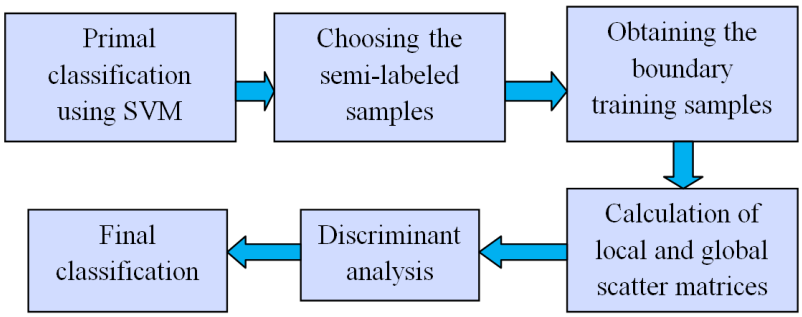

Figure 1. Block diagram of proposed method (HFE-BSL).

These samples are without label before classification. However, their class label is determined after a primal classification. The label of a semi-labeled sample can be correct or incorrect. We add the semi-labeled samples to the original training samples. Then, a useful subset composed of boundary training samples is selected from among a new training set for feature extraction and final classification. The use of this extended training set with relatively high confidence increases the discrimination ability, and so, improves the classification accuracy. To deal with multi-modal data, we combine the local criterion with the global one to provide robustness and flexibility. Using the hybrid criterion, the discriminant analysis works well for more general distributed data. SVM is an appropriate classifier for hyperspectral images [39-42], used effectively in this paper. The novelties of the proposed feature extraction method in this paper are briefly represented as follow:

1) Using SVM as a primal classifier to obtain high confidence semi-labeled samples.
2) Obtaining the boundary training samples.
3) Using a hybrid criterion in the discriminant analysis to obtain both advantages of the local and global criteria.

The remainder of this paper is organized as what follows: The proposed feature extraction method is introduced in section 2. The experimental results are given in section 3. Finally, section 4 concludes the paper.

2. Proposed method

In this section, we introduce the proposed method, hybrid feature extraction based on boundary semi-labeled samples (HFE-BSL). A block diagram of the proposed method is illustrated in Figure 1. We use SVM in three different sections in this paper:

1) SVM as a primal classifier for obtaining high confidence semi-labeled samples.
2) Applying SVM to the new training set (original training plus semi-labeled samples) to obtain SVs and use them as boundary training samples.
3) SVM as a final classifier for obtaining the final classification map (any other classifier can also be used to provide the final classification map).

2.1 Support vector machine

At first, we consider two class linearly separable states. Let $x_i \in \mathbb{R}^d, i = 1, 2, \dots, N_t$ be the training set where d is the number of bands or dimensions of the dataset, and N_t denotes the number of overall training samples. These training samples belong to either of the classes w_1 , and w_2. The corresponding class labels are denoted as $y_i \in \{\mp 1\}$. The goal is to design a hyperplane $g(x) = w^T x + w_0$ that classifies all training sample correctly and leaves the maximum margin from both classes. The following quadratic optimization problem has to be solved:

$$\min J\left(w, w_0\right) = \frac{1}{2}\|w\|^2 \; subject\, to\, y_i\left(w^T x_i + w_0\right) \geq 1, \quad i = 1,2,\dots,N_t \tag{1}$$

In order to solve this optimization problem, the Lagrangian function is defined as:

$$L\left(w, w_0, \lambda\right) = \frac{1}{2} w^T w - \sum_{i=1}^{N_t} \lambda_i \left[y_i\left(w^T x_i + w_0\right) - 1 \right] \tag{2}$$

where $\boldsymbol{\lambda}$ is a vector consisting of the Lagrange multipliers, λ_i. The Lagrange multipliers are zero or positive. Feature vectors (x_i) that are associative with non-zero Lagrange multipliers $(\lambda_i \neq 0)$ are known as support vectors (SVs). SVs lie on either of the two separating hyperplanes: $w^T x + w_0 = \mp 1$. In other words, SVs are a subset of training samples that are located at the boundary between classes. Therefore, SVs can be considered as boundary training samples, and

thus, they are a useful subset of training samples for class discrimination. To allow some training errors for generalization, the slack variables and the associated regularization parameter can be used for non-separable classes. If data cannot be linearly separated, the kernel trick is used to project data into a higher dimensional feature space.

2.2 Production of semi-labeled samples

In the first step of HFE-BSL, we obtain the high confidence semi-labeled samples. Semi-labeled samples are samples whose labels are not known at first and are determined after an initial classification. The collection of reliable training samples is very expensive in terms of time and finance, and it is not common to exploit large ground truth information. To address this issue, the kernel methods such as SVMs have been widely used due to their insensitivity to the curse of dimensionality. They have a high ability to perform with limited training sets, and so they can significantly help in addressing ill-posed problems based on limited training samples. Therefore, in this work, we use SVM for the determination of high confidence semi-labeled samples.

After an initial classification, using original training samples, we randomly choose n_{sl} semi-labeled samples in each class. To have the fair behavior with all classes, we select the same number of semi-labeled samples from all classes. Moreover, to obtain a more classification accuracy, it is appropriate to select much number of semi-labeled samples. To reach this purpose, the maximum number of semi-labeled samples that can be selected is obtained as follows:

$$n_{sl} = \min_{i=1,\ldots,c} n_i \tag{3}$$

where n_i is the number of samples to which the class i label is assigned in the primary classification.

2.3 Obtaining boundary training samples

After production of semi-labeled samples with relatively high confidence in the first step, we add them to the original training samples. Let n_t be the number of original training samples and $n_{ct} = n_{sl} + n_t$ denote the number of new training samples (original training samples plus semi-labeled samples). To increase the classification accuracy, we only use a useful subset of new training samples that play an important role in class discrimination. This useful subset consists of the training samples that are located at the boundary between classes. We train SVM using a new training set, consisting of the original and semi-labeled samples, to obtain SVs from them. In linear SVM, SVs are the samples that are located at the separating hyperplanes: $w^T x + w_0 = \mp 1$. In non-linear SVM, SVs are located within margins where $0 \le y_i(w^T x_i + w_0) < 1$ or in the opposite margins where $y_i(w^T x_i + w_0) < 0$. However, the use of SVs as boundary training samples increases the class discrimination. The obtained training subset, composed of SVs, is used for calculation of between-class and within-class scatter matrices in the discriminant analysis.

2.4 Discriminant analysis with hybrid criterion

After obtaining the boundary training samples, we use them for calculation of global and local scatter matrices as follows:

$$S_w^g = \sum_{k=1}^{c}\left(\sum_{i=1}^{n_k}\left(x_{i,k} - m_k\right)\left(x_{i,k} - m_k\right)^T\right) \tag{4}$$

$$S_b^g = \sum_{k=1}^{c} n_k \left(m_k - m\right)\left(m_k - m\right)^T \tag{5}$$

$$S_w^l = \sum_{k=1}^{c}\left(\sum_{i=1}^{n_k}\left(\sum_{x_+ \in N_{k_1}^+(x_{i,k})}\left(x_{i,k} - x_+\right)\left(x_{i,k} - x_+\right)^T\right)\right) \tag{6}$$

$$S_b^l = \sum_{k=1}^{c}\left(\sum_{i=1}^{n_k}\left(\sum_{x_- \in N_{k_2}^-(x_{i,k})}\left(x_{i,k} - x_-\right)\left(x_{i,k} - x_-\right)^T\right)\right) \tag{7}$$

where, S_w^g, S_b^g, S_w^l, and S_b^l are the global within-class, global between-class, local within-class, and local between-class scatter matrices, respectively. n_k denotes the number of boundary training samples, or SVs, in the kth class, m_k is the mean of SVs in kth class, m is the mean of entire SVs and $\mathrm{x}_{i,k}$ is the ith SV in the kth class. $N_{k_1}^+(\mathrm{x}_{i,k})$ is the set of k_1 nearest neighbours of $\mathrm{x}_{i,k}$ that have the same class label as $\mathrm{x}_{i,k}$, $N_{k_2}^-(\mathrm{x}_{i,k})$ is the set of k_2 nearest neighbours of $\mathrm{x}_{i,k}$ that have different class labels, and x_+ and x_- denote the element of $N_{k_1}^+(\mathrm{x}_{i,k})$ and $N_{k_2}^-(\mathrm{x}_{i,k})$, respectively. We define the hybrid within-class scatter matrix $\left(S_w^h\right)$ and the hybrid between-class scatter matrix $\left(S_b^h\right)$ as follow:

$$S_w^h = \alpha S_w^g + \left(1-\alpha\right) S_w^l \tag{8}$$

$$S_b^h = \alpha S_b^g + \left(1-\alpha\right) S_b^l \tag{9}$$

where $\alpha \in [0,1]$ is a non-negative parameter for providing a trade-off between the global and local terms. $\alpha = 1$ is equivalent to the traditional LDA

(that includes just the global information). Moreover, by considering $\alpha = 0$, just the local information is included in the discriminant analysis. A favorable compromise between robustness and flexibility can be gained by choosing a proper value for the parameter α. An appropriate value for α can be found by searching over the nonnegative values in $[0,1]$. The projection matrix of HFE-BSL is composed of the eigenvectors $S_w^{h\,-1} S_b^h$.

3. Experimental results

We used one synthetic multi-spectral dataset and three real hyperspectral images in our experiments. The objective of the experiments with the synthetic data is assessment and characterization of the proposed method in a fully controlled environment, whereas the objective of the experiments with real datasets is to compare the performance of the proposed method with other methods in the literature. A synthetic image with 80 × 120 pixels was generated. The synthetic scene has eight classes that contain linear mixtures of a set of spectral signatures randomly selected from a digital spectral library compiled by U.S. Geological Survey (USGS), which is available online in "http://speclab.cr.usgs.gov/spectral-lib.html". The USGS library contains spectral plots for nearly 500 materials in the 400-2500 nm spectral range, where the bands have been convolved to the number of bands available for Airborne Visible Infra-Red Imaging Spectrometers (AVIRIS) that comprises 224 spectral bands.

The first real hyperspectral dataset is the Indian pines image that was acquired by AVIRIS. This agriculture image consists of 145×145 pixels and 16 classes, 10 classes of it which were chosen for our experiments. This image comprises 224 spectral bands in the wavelength range from 0.4 to 2.5 μm, nominal spectral resolution of 10 nm, and spatial resolution of 20 m by pixel. The water absorption bands were then deleted, resulting in 200 bands. The second dataset is the University of Pavia that was acquired by Reflective Optics System Imaging Spectrometer (ROSIS). This urban image contains 9 classes and 610×340 pixels, with a spatial resolution of 1.3 m per pixel. The number of spectral bands in the original recorded image is 115 (with a spectral range from 0.43 to 0.86 μm). 12 noisy bands were removed, and the remaining 103 bands were used for the experiments. The Salinas scene is the third dataset used, which was acquired by AVIRIS. It was taken at low altitude with a pixel size of 3.7 m. It consists of 16 classes and 512×217 pixels. It contains 224 spectral bands from 0.4 to 2.5 μm, with a nominal spectral resolution of 10 nm. In this dataset, 20 water absorption bands were removed and 204 bands remained. In this section, we evaluated the efficiency of HFE-BSL in comparison with LDA, SDA, and GDA. We used some measures for the classification evaluation. Accuracy and reliability are defined as follow:

$$ACC_i = \frac{\gamma_i}{A} \tag{10}$$

$$REL_i = \frac{\gamma_i}{B} \tag{11}$$

where, ACC_i and REL_i are the accuracy and reliability of the ith class, respectively. γ_i is the number of testing samples that are correctly classified, A is the total testing samples of class i, and B is the total samples labeled as class i. We used the average accuracy (AA), overall accuracy (OA), and average reliability (AR) for evaluation of the classification accuracy, which are calculated as follow:

$$AA = \frac{1}{c}\sum_{i=1}^{c} ACC_i \tag{12}$$

$$AR = \frac{1}{c}\sum_{i=1}^{c} REL_i \tag{13}$$

$$OA = \frac{\sum_{i=1}^{c} \gamma_i}{D} \tag{14}$$

where, D is the total number of testing samples. We used four classifiers to obtain the final classification maps: SVM, ML, 1-nearest-neighbor (1NN), and distance-weighted discrimination (DWD) [7], [10]. We implemented SVM with the radial basis function (RBF) kernel. The use of RBF kernel is a common choice, because it has less numerical difficulties, and it is easy to be tuned. In addition, the RBF kernel is a universal kernel and includes the other valid kernels as particular cases. We used the Library for Support Vector Machines (LIBSVM) tool for the implementation of SVM [43]. The one-against-one multi-class classification algorithm was used in the experiments. The SVM parameters were set as follow: the penalty parameter C was tested between $[10-1000]$ with a step size increment of 20, and γ parameter of the RBF kernel was tested between $[0.1-2]$ with a step size increment of 0.1. The best values were obtained using a 5-fold cross-validation approach.

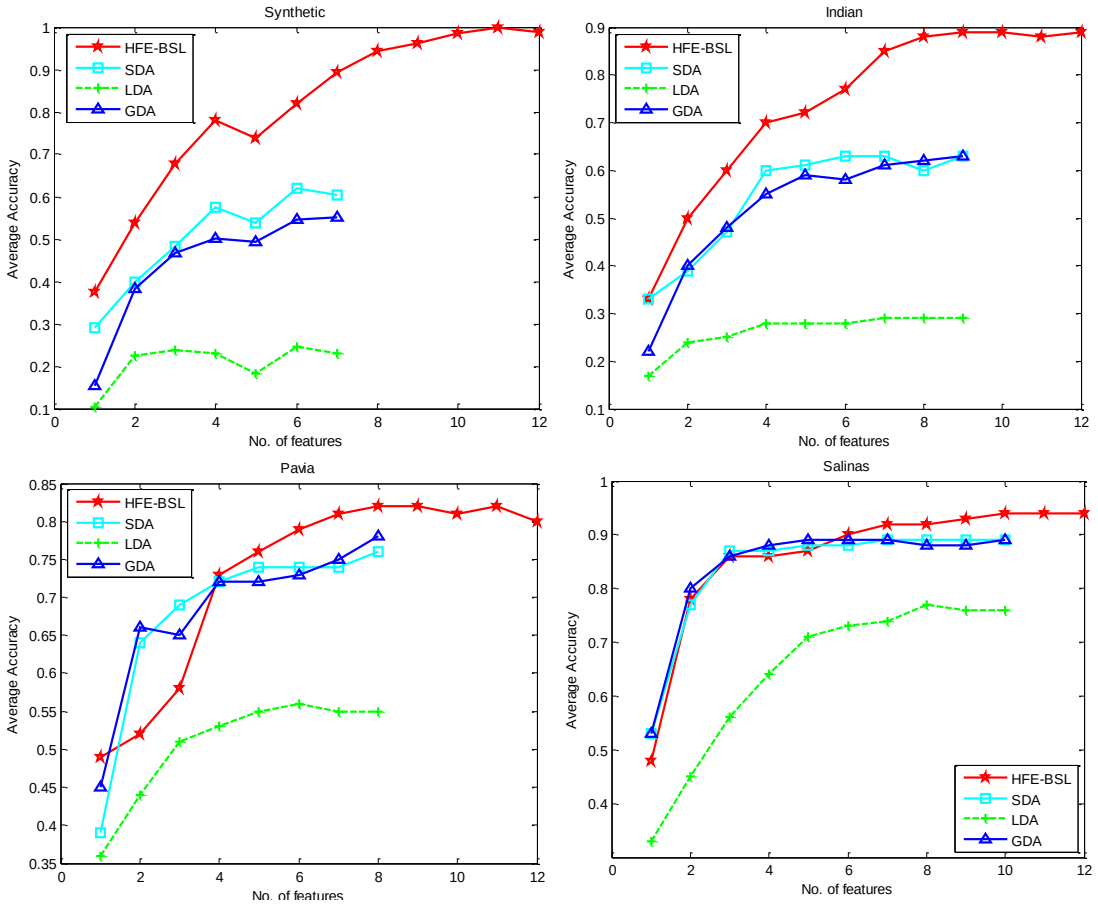

Figure 2. Classification accuracy versus number of extracted features using SVM classifier and 16 training samples.

The training samples were chosen randomly from the entire of the scene, and the remaining samples were used as the testing data. Due to the random selection of training samples, each experiment was repeated 10 times, and the average results were reported. The same training samples were used for all methods to obtain a fair comparison. We considered $\alpha = 0.5$ in our experiments to have both the robustness and flexibility as the same. The classification accuracies versus the number of extracted features, obtained by the SVM classifier and 16 original training samples, for synthetic, Indian, Pavia and Salinas datasets are shown in Figure 2. The accuracies and reliabilities for each class of Indian dataset obtained by the SVM classifier, 9 extracted features and 16 original training samples are represented in Table 1. Moreover, the average accuracy, average reliability, overall accuracy, and execution time for all the compared methods are reported in this table. This type of table, obtained by 16 original training samples, is given for Pavia with 8 extracted features (see Table 2) and for Salinas with 10 extracted features (see Table 3). The comparison of execution times, which are reported in Tables 1-3, show that the proposed method has more computation time than the other feature extraction methods. This is expected, because the proposed method uses the SVM classifier at least two times: the first time, to obtain the initial classification map for providing the semi-labeled samples, and the second time, to obtain SVs as boundary samples. The ground truth map (GTM) and the classification maps of synthetic (with 7 extracted features), Indian (with 9 extracted features), Pavia (with 8 extracted features), and Salinas (with 10 extracted features) are shown in Figures 3-6, respectively.

The classification results obtained using different classifiers (SVM, ML, 1NN, DWD) with feature extracted by HFE-BSL, LDA, SDA, GDA and also using the original features (without feature extraction) are reported in Table 4. This table provides the highest average classification accuracies achieved by four classifiers and 16 original training samples. The numbers in the parentheses represent the number of features achieving the highest average accuracies in the experiments. For example the entry in the first column and the second row of Table 4, 92.34 ± 1.08 (6), means that the highest classification accuracy, which is obtained by the HFE-BSL feature extraction method and ML classifier, is provided by 6 extracted features and is equal to 92.34 with a standard deviation of 1.08.

Table 1. Accuracies and reliabilities obtained for each class of Indian dataset using 9 extracted features and 16 training samples.

Class			HFE-BSL		SDA		LDA		GDA	
No	Name of class	# samples	Acc.	Rel.	Acc.	Rel.	Acc.	Rel.	Acc.	Rel.
1	Corn-no till	1434	0.83	0.77	0.56	0.45	0.19	0.30	0.62	0.49
2	Corn-min till	834	0.85	0.60	0.41	0.48	0.33	0.16	0.39	0.44
3	Grass/pasture	497	0.99	0.93	0.89	0.49	0.29	0.39	0.73	0.45
4	Grass/trees	747	0.99	0.98	0.68	0.77	0.36	0.18	0.74	0.80
5	Hay-windrowed	489	1.00	1.00	0.92	0.97	0.49	0.74	0.98	1.00
6	Soybeans-no till	968	0.82	0.69	0.62	0.58	0.34	0.17	0.63	0.43
7	Soybeans-min till	2468	0.61	0.89	0.60	0.69	0.08	0.40	0.49	0.75
8	Soybeans-clean till	614	0.92	0.76	0.42	0.48	0.23	0.11	0.45	0.46
9	Woods	1294	0.92	1.00	0.80	0.92	0.25	0.60	0.70	0.86
10	Bldg-Grass-Tree-Drives	380	1.00	0.80	0.41	0.44	0.31	0.18	0.58	0.44
Average Accuracy and Average Reliability			0.89	0.84	0.63	0.63	0.29	0.32	0.63	0.61
Overall Accuracy			0.92		0.83		0.61		0.82	
Execution Time (s)			42.71		37.60		0.86		12.01	

Table 2. Accuracies and reliabilities obtained for each class of Pavia dataset using 8 extracted features and 16 training samples.

Class			HFE-BSL		SDA		LDA		GDA	
No	Name of class	# samples	Acc.	Rel.	Acc.	Rel.	Acc.	Rel.	Acc.	Rel.
1	Asphalt	6631	0.83	0.89	0.70	0.88	0.25	0.63	0.69	0.90
2	Meadows	18649	0.71	0.90	0.69	0.84	0.52	0.87	0.61	0.87
3	Gravel	2099	0.71	0.57	0.61	0.46	0.44	0.26	0.72	0.50
4	Trees	3064	0.95	0.82	0.70	0.59	0.80	0.62	0.85	0.65
5	Painted metal sheets	1345	1.00	1.00	0.99	0.92	0.96	1.00	0.99	0.98
6	Bare Soil	5029	0.74	0.43	0.61	0.41	0.72	0.25	0.67	0.36
7	Bitumen	1330	0.85	0.60	0.93	0.49	0.29	0.17	0.89	0.46
8	Self-Blocking Bricks	3682	0.61	0.73	0.64	0.70	0.24	0.40	0.62	0.69
9	Shadows	947	1.00	1.00	1.00	1.00	0.71	0.61	1.00	0.97
Average Accuracy and Average Reliability			0.82	0.77	0.76	0.70	0.55	0.53	0.78	0.71
Overall Accuracy			0.89		0.80		0.73		0.82	
Execution Time (s)			57.94		46.35		1.53		16.83	

Table 3. Accuracies and reliabilities obtained for each class of Salinas using 10 extracted features and 16 training samples.

Class			HFE-BSL		SDA		LDA		GDA	
No	Name of class	# samples	Acc.	Rel.	Acc.	Rel.	Acc.	Rel.	Acc.	Rel.
1	Brocoli_green_weeds_1	2009	1.00	1.00	0.94	1.00	0.96	0.98	0.90	0.99
2	Brocoli_green_weeds_2	3726	1.00	1.00	0.99	0.96	0.95	0.98	0.98	0.93
3	Fallow	1976	1.00	1.00	0.86	0.85	0.63	0.62	0.96	0.87
4	Fallow_rough_plow	1394	1.00	0.99	0.99	0.98	0.81	0.89	1.00	0.85
5	Fallow_smooth	2678	0.99	1.00	0.93	0.90	0.68	0.76	0.96	0.97
6	Stubble	3959	1.00	1.00	0.96	1.00	0.91	1.00	0.97	1.00
7	Celery	3579	1.00	1.00	0.99	0.97	0.95	1.00	0.99	0.94
8	Grapes_untrained	11271	0.53	0.62	0.54	0.73	0.50	0.68	0.52	0.70
9	Soil_vineyard_develop	6203	0.98	0.99	0.96	0.98	0.65	0.94	0.99	0.99
10	Corn_senesced_green_weeds	3278	0.98	0.96	0.79	0.81	0.82	0.38	0.79	0.89
11	Lettuce_romaine_4weeks	1068	1.00	1.00	0.86	0.80	0.80	0.94	0.84	0.96
12	Lettuce_romaine_5 weeks	1927	1.00	0.99	0.99	0.82	0.56	0.54	0.99	0.91
13	Lettuce_romaine_6 weeks	916	1.00	1.00	1.00	0.67	0.72	0.48	1.00	0.74
14	Lettuce_romaine_7 weeks	1070	1.00	1.00	0.84	0.92	0.68	0.72	0.84	0.58
15	Vineyard_untrained	7268	0.49	0.40	0.68	0.51	0.58	0.46	0.66	0.51
16	Vineyard_vertical_trellis	1807	1.00	1.00	0.90	0.98	0.91	1.00	0.82	0.95
Average Accuracy and Average Reliability			0.94	0.93	0.89	0.87	0.76	0.77	0.89	0.86
Overall Accuracy			0.95		0.91		0.84		0.91	
Execution Time (s)			73.66		51.11		3.05		21.71	

In addition to SDA, which is a popular semi-supervised feature extraction method, we compared our semi-supervised proposed method with two state-of-the-art semi-supervised feature extraction methods: semi-supervised local discriminant analysis (SELD) [45] and semi-supervised probabilistic principal component analysis (S^2PPCA) [46]. The aim of SELD is to find a projection for preserving the local neighborhood information and maximizing the class discrimination of the data. In the SELD method, an unsupervised method (from the class of local linear feature extraction methods such as neighborhood preserving embedding (NPE)) and a supervised method (LDA) are combined without any tuning parameters.

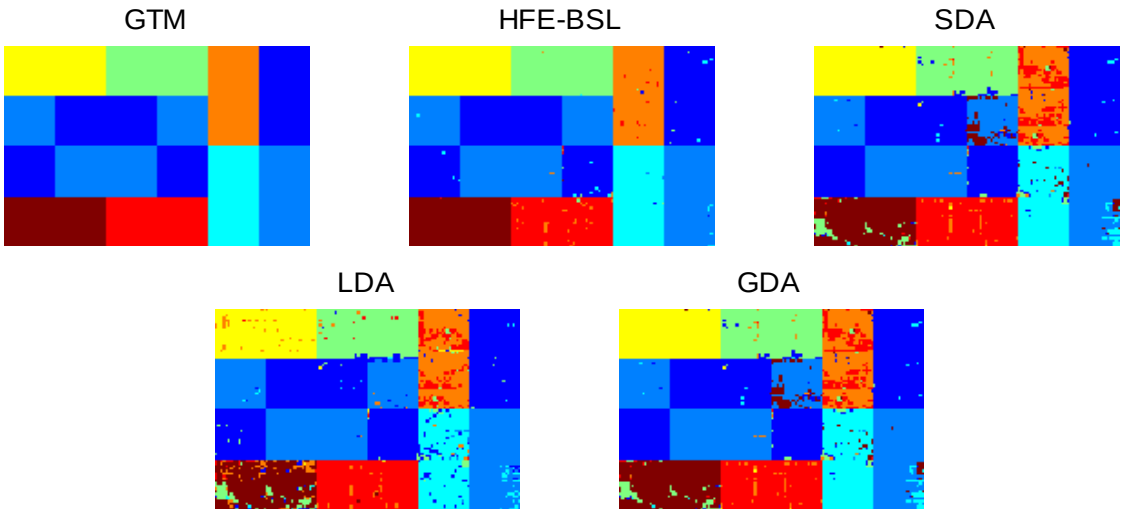

Figure 3. GTM and classification maps for synthetic dataset (16 training samples and 7 extracted features are used).

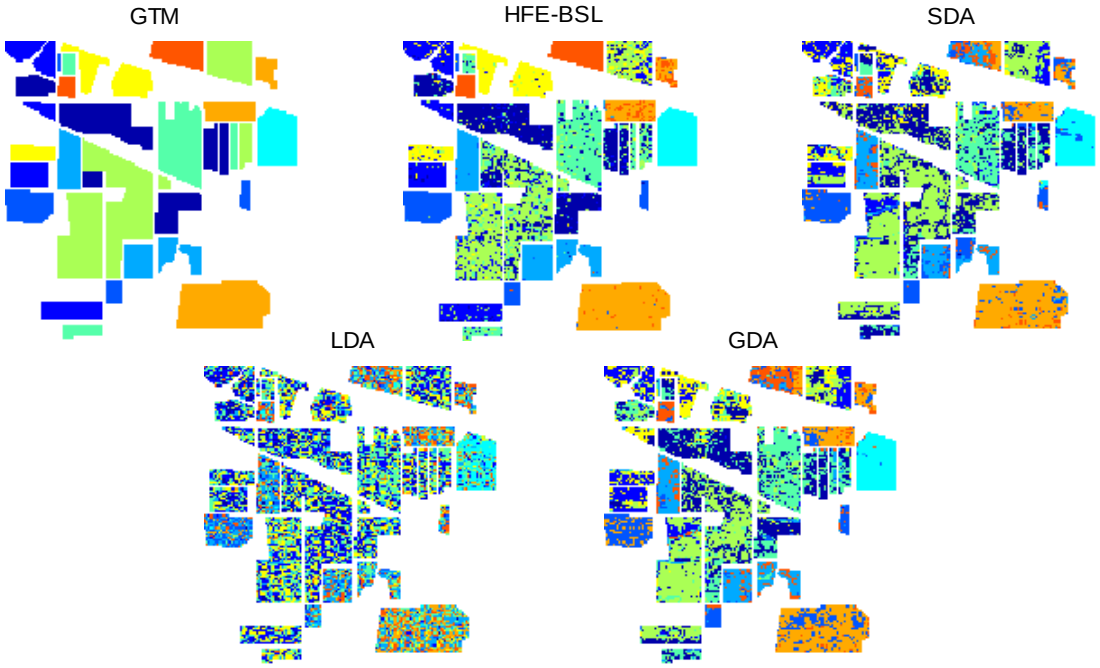

Figure 4. GTM and classification maps for Indian dataset (16 training samples and 9 extracted features are used).

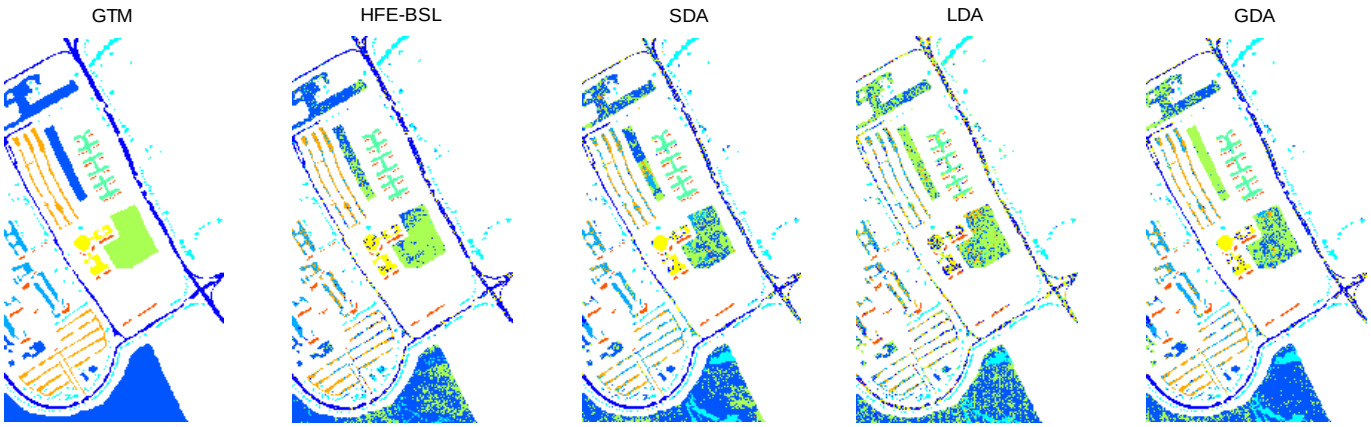

Figure 5. GTM and classification maps for Pavia dataset (16 training samples and 8 extracted features are used).

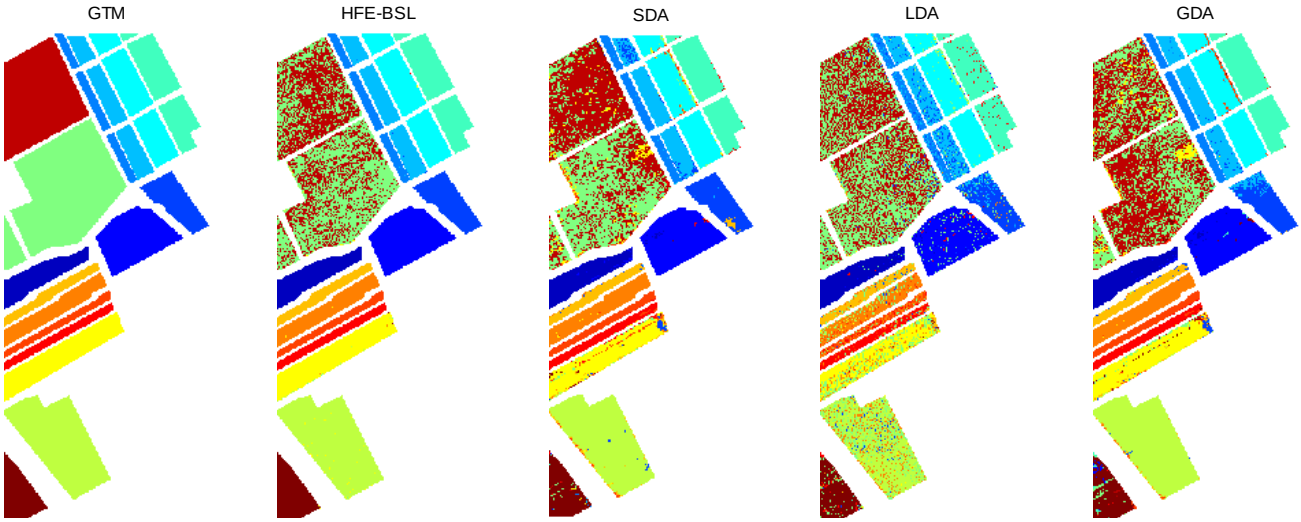

Figure 6. GTM and classification maps for Salinas dataset (16 training samples and 10 extracted features are used).

Table 4. Mean and standard deviation of highest average classification accuracies (numbers in parentheses represent number of features achieving highest average accuracies in experiments).

Dataset	Classifier	HFE-BSL	SDA	LDA	GDA	Original features
Synthetic	SVM	94.21±0.43 (7)	83.34±0.76 (6)	58.12±2.01 (7)	82.13±2.52 (5)	84.32±0.76 (224)
	ML	92.34±1.08 (6)	79.12±1.52 (4)	58.03±3.01 (4)	80.09±2.67 (5)	---
	1NN	93.12±0.98 (6)	80.91±0.92 (7)	59.02±1.05 (6)	81.72±2.60 (6)	84.05±1.07 (224)
	DWD	94.11±0.86 (7)	83.78±0.59 (7)	59.15±2.23 (7)	82.01±1.93 (7)	85.21±0.85 (224)
Indian	SVM	89.40±1.14 (9)	63.12±1.02 (6)	28.92±2.45 (7)	63.15±2.90 (9)	65.23±0.63 (200)
	ML	85.37±1.87 (6)	61.83±2.14 (5)	21.36±3.51 (4)	65.11±2.82 (6)	---
	1NN	87.26±1.51 (7)	64.44±1.18 (5)	25.33±2.38 (8)	63.72±2.45 (6)	61.08±3.10 (200)
	DWD	89.53±1.32 (8)	62.97±1.21 (5)	33.15±2.57 (9)	63.13±2.15 (9)	64.75±0.75 (200)
Pavia	SVM	82.39±0.78 (8)	76.29±0.81 (8)	55.65±1.98 (6)	78.38±2.06 (8)	76.13±0.76 (103)
	ML	79.21±2.01 (4)	78.20±1.93 (4)	55.41±2.61 (5)	73.64±1.64 (6)	---
	1NN	81.04±1.52 (6)	72.79±1.93 (7)	57.43±2.14 (8)	71.17±3.11 (4)	71.49±1.72 (103)
	DWD	81.22±1.49 (5)	74.54±1.03 (8)	59.25±1.86 (7)	78.68±2.01 (7)	77.11±1.13 (103)
Salinas	SVM	93.76±1.13 (10)	89.12±0.77 (7)	76.83±1.98 (8)	89.02±2.43 (5)	87.26±0.84 (204)
	ML	88.47±1.46 (6)	89.61±2.35 (9)	76.13±2.02 (5)	87.02±1.69 (4)	---
	1NN	91.45±1.24 (5)	89.28±1.88 (5)	77.03±1.76 (6)	89.88±1.94 (12)	86.36±2.03 (204)
	DWD	92.45±1.31 (4)	90.03±1.52 (9)	77.74±1.06 (10)	90.45±1.62 (4)	88.43±0.91 (204)

The supervised and unsupervised methods are not linearly combined. In other words, instead of using both the labeled and unlabeled samples together, at first, the samples were divided into two sets: labeled and unlabeled. Then, the labeled samples were employed through the supervised method only, and the unlabeled ones were employed through the unsupervised, locality preserving method only. Thus, the local neighborhood information is preserved inferred from unlabeled samples, while the class discrimination of the data is maximized inferred from the labeled samples. The supervised probabilistic principal component analysis (SPPCA) is the extension of the probabilistic PCA (PPCA) which incorporates the label information into the projection. In addition to the inter-covariance between the inputs and outputs, SPPCA takes into account the intra-covariance of both. S^2PPCA incorporates both the labeled and

unlabeled information into the projections. Both SPPCA and S^{2}PPCA use an expectation maximization algorithm to generate the mapping. The experimental results for these methods compared to the others are reported in Table 4.

Moreover, we used the McNemars test to assess the statistical significance of differences in the classification results. The McNemars test is based upon the standardized normal test statistic [44]:

$$Z_{12} = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}} \quad (15)$$

The number of samples correctly by classifier 1 and incorrectly by classifier 2 is denoted by f_{12}. The difference in the accuracy between the classifiers 1 and 2 is said to be statistically significant if $|Z_{12}| > 1.96$. The sign of Z_{12} indicates whether classifier 1 is more accurate than classifier 2 ($Z_{12} > 0$) or vice versa ($Z_{12} < 0$). The results of the McNemars tests for different cases with 16 original training samples are shown in Table 5.

For example, in table (Indian/SVM classifier/9 extracted features), $|Z_{14}| > 0$, means that HFE-BSL is superior to GDA and also because $|Z_{14}| = 37.23 > 1.96$, this difference is significant from the statistical view point.

Because of the singularity of within-class scatter matrix in small sample size situation, LDA had the worst efficiency. The use of kernel trick and the use of semi-supervised approach improved the classification accuracy of GDA and SDA, respectively. One sees from the results obtained that HFE-BSL can provide the highest classification accuracy compared to the other feature extraction methods. Note that the rank of between-class scatter matrix (S_b) is limited in the LDA, GDA, and SDA methods. Thus, these methods can be extract maximum $c-1$ features. In the proposed method, HFE-BSL, local between-class scatter matrix (S_b^l) has a non-parametric form, and thus, its rank is not limited to the number of classes. Therefore, the rank of the hybrid between-class scatter matrix (S_b^h) is also not limited. Thus, the HFE-BSL method can extract more than $c-1$ features. In what following, we represent the classification accuracies by SVM, 1NN, and ML trained by original training samples as a baseline. When we use only the original training samples (without any semi-labeled samples) in the proposed method, we name it hybrid feature extraction (HFE). We assessed the effect of training sample size on the classification accuracy. The classification results were obtained using $n_t = 5, 10, 15, 20,$ where n_t is the number of original available training samples. The classification accuracies achieved by different numbers of original training samples, different feature extraction methods, and different classifiers are reported in Tables 6, 7, 8, and 9 for the synthetic, Indian, Pavia, and Salinas datasets respectively.

The most important results can be reviewed as follow:

- The HFE-BSL method provided better classification results compared to the LDA, GDA, SDA, SELD, and S^{2}PPCA methods in terms of classification accuracy (average accuracy, average reliability, and overall accuracy).
- The higher classification accuracy of HFE-BSL compared to other methods was significant from the statistical view point.
- More accurate classification maps (with less noise) were obtained using HFE-BSL.
- The efficiency of the HFE-BSL method was assessed using 4 different classifiers: SVM (a kernel-based learning machine with low sensitivity to the training sample size), ML (a parametric classifier with high sensitivity to the training sample size), NN (a simple nonparametric classifier), and DWD (which avoids data piling, and can give improved generalizability).
- The better performance of HFE-BSL compared to other methods is shown in different numbers of extracted feature and also in different numbers of training samples.

The main advantages of the proposed method can be represented as follow: HFE-BSL uses the ability of SVM, which is a state-of-the-art classifier for high dimensional data with low sensitivity to the training set size, to obtain the semi-labeled samples with high confidence. HFE-BSL uses a rich and useful subset of semi-labeled samples, i.e., boundary samples, to increase the class discrimination. Moreover, the HFE-BSL method can simultaneously provide robustness and flexibility to deal with multi-modal data using both the global and local information for estimation of scatter matrices. The disadvantage of HFE-BSL is its more computation time than the competitor methods such as LDA, GDA, and SDA. This increased elapsed time is due to using SVM at least two times (the first time to obtain the initial classification map for providing the semi-labeled samples, and the second one to obtain SVs as boundary samples).

Table 5. Statistical significance of differences in classification (Z). Each case in table represents Z_{rc}, where r is row and c is column (all tables obtained by 16 training samples).

Indian/SVM classifier/9 extracted features				
	HFE-BSL	SDA	LDA	GDA
HFE-BSL	0	33.50	73.27	37.23
SDA	-33.50	0	52.51	4.72
LDA	-73.27	-52.51	0	-50.13
GDA	-37.23	-4.72	50.13	0

Indian/1NN classifier/7 extracted features				
	HFE-BSL	SDA	LDA	GDA
HFE-BSL	0	9.23	50.52	9.14
SDA	-9.23	0	45.06	1.62
LDA	-50.52	-45.06	0	-43.37
GDA	-9.14	-1.62	43.37	0

Pavia/SVM classifier/8 extracted features				
	HFE-BSL	SDA	LDA	GDA
HFE-BSL	0	2.12	50.51	6.96
SDA	-2.12	0	52.06	8.53
LDA	-50.51	-52.06	0	-46.80
GDA	-6.96	-8.53	46.80	0

Pavia/1NN classifier/6 extracted features				
	HFE-BSL	SDA	LDA	GDA
HFE-BSL	0	14.68	59.05	15.78
SDA	-14.68	0	51.92	2.82
LDA	-59.05	-51.92	0	-48.08
GDA	-15.78	-2.82	48.08	0

Salinas/SVM classifier/10 extracted features				
	HFE-BSL	SDA	LDA	GDA
HFE-BSL	0	6.22	46.25	0.27
SDA	-6.22	0	42.14	-5.47
LDA	-46.25	-42.14	0	-45.60
GDA	-0.27	5.47	45.60	0

Salinas/1NN classifier/5 extracted features				
	HFE-BSL	SDA	LDA	GDA
HFE-BSL	0	14.93	55.41	19.03
SDA	-14.93	0	46.12	4.06
LDA	-55.41	-46.12	0	-43.53
GDA	-19.03	-4.06	43.53	0

Indian/ML classifier/6 extracted features				
	HFE-BSL	SDA	LDA	GDA
HFE-BSL	0	19.25	55.24	12.41
SDA	-19.25	0	43.17	-5.80
LDA	-55.24	-43.17	0	-46.97
GDA	-12.41	5.80	46.97	0

Indian/DWD classifier/8extracted features				
	HFE-BSL	SDA	LDA	GDA
HFE-BSL	0	44.65	71.30	43.17
SDA	-44.65	0	41.79	-1.06
LDA	-71.30	-41.79	0	-41.70
GDA	-43.17	1.06	41.70	0

Pavia/ML classifier/4 extracted features				
	HFE-BSL	SDA	LDA	GDA
HFE-BSL	0	6.71	57.59	20.10
SDA	-6.71	0	54.35	14.09
LDA	-57.59	-54.35	0	-45.72
GDA	-20.10	-14.09	45.72	0

Pavia/DWD classifier/5 extracted features				
	HFE-BSL	SDA	LDA	GDA
HFE-BSL	0	56.28	106.58	60.73
SDA	-56.28	0	60.71	8.36
LDA	-106.58	-60.71	0	-53.18
GDA	-60.73	-8.36	53.18	0

Salinas/ML classifier/6 extracted features				
	HFE-BSL	SDA	LDA	GDA
HFE-BSL	0	0.53	53.31	1.91
SDA	-0.53	0	37.45	-20.98
LDA	-53.31	-37.45	0	-56.59
GDA	-1.91	20.98	56.59	0

Salinas/DWD classifier/4 extracted features				
	HFE-BSL	SDA	LDA	GDA
HFE-BSL	0	45.18	73.63	45.51
SDA	-45.18	0	44.32	0.42
LDA	-73.63	-44.32	0	-44.19
GDA	-45.51	-0.42	44.19	0

4. Conclusion

We proposed a new feature extraction method in this paper. This method, called HFE-BSL, uses the ability of high confidence semi-labeled samples to cope with the small sample size problems. HFE-BSL uses a useful subset of training samples composed of boundary training samples for increasing the classification accuracy. The proposed feature extraction method combines both the global and local criteria for calculation of scatter matrices in discriminant analysis.

Thus, HFE-BSL is robust and flexible. The experiments carried out using four hyperspectral images indicated the good efficiency of the proposed approach in comparison with some popular and state-of-the-art feature extraction methods in small sample size situations.

The proposed method just uses the rich spectral information and does not consider the valuable spatial information. In the future works, we will try to use the spatial information contained in a neighborhood window to increase the reliability and accuracy of the semi-labeled samples. It is expected that the use of both the spectral and spatial information improves the classification performance.

Table 6. Highest average classification accuracies and their corresponding standard deviations achieved using SVM, ML, 1NN and DWD classifiers with features extracted by HFE-BSL, HFE, SDA, LDA, GDA, and using original features for synthetic dataset. Number in parenthesis represents number of features achieving highest accuracies in experiments.

Feature extraction	Classifier	n_t (the number of original training samples)			
		$n_t = 5$	$n_t = 10$	$n_t = 15$	$n_t = 20$
HFE-BSL	SVM	91.12±0.47(7)	93.31±0.49(7)	94.78±1.05(6)	95.12±0.97(7)
	ML	86.05±1.13(5)	89.54±1.09(5)	90.87±1.67(4)	91.92±1.83(6)
	1NN	89.35±0.98(6)	92.68±0.88(8)	93.38±0.90(9)	93.84±0.64(6)
	DWD	92.13±1.21(8)	93.22±1.32(7)	93.75±1.99(7)	94.98±1.36(7)
HFE	SVM	80.45±0.54(4)	86.46±0.72(6)	90.54±1.01(7)	93.53±1.13(7)
	ML	78.87±1.87(5)	83.65±2.08(5)	89.76±2.54(6)	90.64±1.05(6)
	1NN	80.36±0.99(7)	85.74±1.62(7)	90.34±1.43(7)	92.87±0.98(7)
	DWD	81.07±1.45(7)	86.33±0.86(6)	90.67±1.53(7)	92.54±2.32(7)
SDA	SVM	82.43±1.32(5)	86.18±0.53(5)	89.34±1.02(6)	91.20±1.83(4)
	ML	80.64±1.09(6)	84.56±0.1.76(7)	87.76±1.50(4)	90.67±1.09(4)
	1NN	82.76±2.07(7)	85.76±2.31(6)	88.98±1.09(5)	90.86±1.32(6)
	DWD	83.11±1.98(7)	85.85±1.96(6)	89.54±1.11(7)	91.09±0.98(7)
LDA	SVM	62.13±2.43(6)	74.21±2.11(7)	78.65±1.14(6)	85.52±2.15(7)
	ML	55.09±3.07(7)	68.64±3.011(6)	72.64±2.01(7)	81.98±2.64(7)
	1NN	57.87±2.87(6)	66.38±2.09(6)	73.87±3.73(6)	83.98±3.03(6)
	DWD	61.09±3.08(70	73.65±3.43(7)	77.98±2.52(6)	84.87±2.65(5)
GDA	SVM	80.32±1.98(7)	86.18±1.16(6)	89.14±2.06(5)	90.67±0.86(5)
	ML	79.04±0.98(5)	86.24±2.03(7)	88.24±3.04(6)	89.19±0.98(4)
	1NN	81.09±2.06(6)	87.76±1.07(6)	89.32±2.09(6)	89.02±1.09(6)
	DWD	81.63±1.84(6)	87.02±1.36(7)	89.01±2.76(5)	90.28±0.99(7)
Original features	SVM	82.14±1.94(224)	87.68±1.15(224)	88.02±0.65(224)	90.16±1.53(224)
	ML	---	---	---	---
	1NN	82.35±2.09(224)	87.98±2.35(224)	89.42±1.63(224)	88.75±1.67(224)
	DWD	82.76±2.15(224)	88.87±1.95(224)	89.33±1.23(224)	90.34±1.97(224)

Table 7. Highest average classification accuracies and their corresponding standard deviations achieved using SVM, ML, 1NN and DWD classifiers with features extracted by HFE-BSL, HFE, SDA, LDA, GDA, and using original features for Indian dataset. Number in parenthesis represents number of features achieving highest accuracies in experiments.

Feature extraction	Classifier	n_t (the number of original training samples)			
		$n_t = 5$	$n_t = 10$	$n_t = 15$	$n_t = 20$
HFE-BSL	SVM	86.43±0.42(10)	89.12±0.51(7)	89.11±1.13(8)	89.47±1.01(11)
	ML	79.74±1.34(5)	81.25±1.23(5)	85.20±1.72(6)	84.17±2.05(6)
	1NN	82.29±0.98(7)	84.57±1.49(6)	86.94±1.64(12)	87.35±2.17(5)
	DWD	87.03±0.65(6)	88.97±0.49(6)	89.23±2.01(9)	89.41±0.93(8)
HFE	SVM	48.34±0.76(4)	57.61±0.64(6)	63.34±1.04(6)	73.87±1.06(8)
	ML	44.67±1.97(3)	51.28±1.42(5)	60.81±1.21(6)	71.13±1.29(5)
	1NN	46.23±2.03(7)	52.09±0.75(6)	62.43±0.99(8)	72.74±0.63(5)
	DWD	49.23±0.76(8)	58.04±0.74(6)	62.97±1.41(9)	74.09±0.89(6)
SDA	SVM	58.37±1.08(5)	64.43±1.31(6)	69.75±0.78(6)	73.66±1.41(5)
	ML	51.73±3.02(6)	57.86±2.34(5)	61.81±2.02(4)	69.76±1.76(6)
	1NN	56.22±2.42(8)	63.03±1.98(7)	67.38±1.07(6)	72.79±1.48(7)
	DWD	58.24±0.69(6)	65.11±1.24(6)	70.01±2.01(9)	74.04±0.91(6)
LDA	SVM	23.34±1.02(4)	24.94±1.03(6)	26.69±1.76(7)	26.76±1.53(6)
	ML	22.15±3.44(3)	23.36±2.13(5)	19.05±4.01(5)	25.87±1.98(5)
	1NN	23.11±2.32(7)	23.96±1.67(7)	24.67±2.54(8)	26.13±2.64(5)
	DWD	25.14±.098(6)	24.89±0.94(6)	26.50±2.01(7)	26.37±1.03(7)
GDA	SVM	48.73±0.75(8)	55.29±2.25(9)	62.33±1.11(9)	65.27±2.32(9)
	ML	45.55±2.66(6)	53.93±1.76(4)	65.09±2.92(5)	62.36±3.51(6)
	1NN	47.38±2.45(5)	55.19±3.74(6)	63.14±2.54(6)	64.87±1.65(5)
	DWD	47.87±1.45(7)	55.65±1.72(6)	64.34±1.77(7)	64.97±1.39(7)
Original features	SVM	59.49±0.56(200)	61.69±1.52(200)	65.18±0.71(200)	70.13±0.92(200)
	ML	---	---	---	---
	1NN	51.33±2.36(200)	55.49±1.68(200)	60.68±2.93(200)	68.48±1.68(200)
	DWD	60.02±0.87(200)	60.94±1.42(200)	65.83±0.92(200)	70.32±0.98(200)

Table 8. Highest average classification accuracies and their corresponding standard deviations for Pavia dataset. Number in parenthesis represents number of features achieving highest accuracies in experiments.

Feature extraction	Classifier	n_t (the number of original training samples)			
		$n_t = 5$	$n_t = 10$	$n_t = 15$	$n_t = 20$
HFE-BSL	SVM	79.31±0.52(7)	80.71±0.14(6)	82.28±0.69(8)	82.34±1.11(10)
	ML	72.36±1.20(5)	76.04±0.72(4)	78.94±1.93(4)	79.12±0.98(4)
	1NN	78.62±0.67(7)	79.55±0.49(8)	80.76±1.38(6)	81.02±1.05(6)
	DWD	80.02±0.86(8)	80.54±0.34(7)	81.98±1.03(7)	82.62±1.04(9)
HFE	SVM	55.21±2.42(4)	71.31±0.79(5)	73.23±0.67(7)	75.19±1.38(8)
	ML	48.49±3.01(4)	64.36±1.83(5)	67.74±1.53(6)	70.45±1.90(3)
	1NN	50.72±2.37(6)	68.48±2.25(7)	71.64±2.16(8)	73.62±2.65(7)
	DWD	56.76±1.74(8)	71.01±2.02(8)	72.84±1.46(7)	75.24±1.36(8)
SDA	SVM	71.03±1.41(8)	74.32±0.83(8)	76.02±0.93(7)	77.18±0.48(6)
	ML	68.84±2.06(6)	72.57±0.1.57(7)	74.86±2.01(5)	78.95±0.94(5)
	1NN	68.42±1.89(7)	70.03±1.03(6)	71.90±1.92(7)	76.77±1.05(7)
	DWD	71.34±1.72(7)	73.95±1.03(6)	76.15±1.09(8)	77.13±0.68(7)
LDA	SVM	34.09±2.08(5)	36.04±1.09(6)	52.53±2.10(7)	56.44±1.98(7)
	ML	25.91±2.22(3)	28.48±2.28(5)	51.96±2.52(5)	53.72±3.64(4)
	1NN	27.38±1.54(7)	31.41±1.37(6)	55.23±1.97(8)	54.16±2.46(8)
	DWD	36.33±2.02(7)	36.92±2.01(5)	54.09±1.72(7)	55.65±2.21(6)
GDA	SVM	70.06±1.93(7)	73.23±1.36(6)	78.02±1.93(8)	81.68±1.21(7)
	ML	65.17±2.17(5)	68.10±3.07(3)	72.85±1.54(5)	77.41±1.63(6)
	1NN	66.92±0.89(5)	69.26±2.96(4)	71.09±2.99(4)	76.08±0.99(4)
	DWD	70.32±2.01(6)	73.32±1.35(6)	77.98±1.02(5)	82.01±1.72(5)
Original features	SVM	68.11±0.76(103)	71.48±0.82(103)	74.01±0.69(103)	76.83±0.56(103)
	ML	---	---	---	---
	1NN	67.73±1.19(103)	69.34±1.56(103)	70.31±1.62(103)	71.61±1.03(103)
	DWD	68.03±0.82(103)	72.34±0.72(103)	73.98±1.02(103)	76.74±0.88(103)

Table 9. Highest average classification accuracies and their corresponding standard deviations for Salinas dataset. Number in parenthesis represents number of features achieving highest accuracies in experiments.

Feature extraction	Classifier	n_t (the number of original training samples)			
		$n_t = 5$	$n_t = 10$	$n_t = 15$	$n_t = 20$
HFE-BSL	SVM	90.34±0.25(12)	92.26±0.52(11)	93.69±1.07(10)	93.32±1.03(13)
	ML	82.25±0.76(9)	84.47±1.95(7)	87.74±1.44(7)	88.41±2.08(8)
	1NN	88.84±0.49(8)	89.90±0.64(6)	90.87±1.37(5)	91.27±1.21(4)
	DWD	91.03±0.43(11)	91.34±0.67(10)	92.62±1.72(8)	92.74±0.99(12)
HFE	SVM	76.67±0.23(5)	83.12±0.47(10)	88.49±1.11(9)	92.53±0.97(9)
	ML	70.67±0.68(5)	78.25±1.83(9)	84.05±1.36(7)	86.67±2.25(10)
	1NN	73.24±0.52(6)	80.83±0.72(9)	86.67±1.48(8)	90.47±1.08(7)
	DWD	77.54±0.88(7)	83.01±0.52(10)	87.76±1.05(9)	91.98±0.43(6)
SDA	SVM	83.66±1.08(7)	87.58±0.39(7)	89.01±0.65(6)	92.48±0.97(5)
	ML	79.74±2.08(13)	85.41±2.05(12)	88.98±2.46(9)	89.99±1.12(10)
	1NN	80.25±0.89(8)	84.12±1.82(9)	88.69±2.13(4)	89.24±1.32(7)
	DWD	83.21±0.97(10)	86.23±2.04(10)	89.32±0.76(5)	92.83±0.76(8)
LDA	SVM	60.04±2.03(6)	72.33±1.96(9)	76.82±1.94(8)	84.55±2.05(7)
	ML	58.17±3.05(6)	69.47±2.44(7)	75.23±1.79(5)	83.62±3.46(9)
	1NN	62.27±1.97(6)	73.08±1.99(6)	77.00±1.63(6)	85.28±2.64(7)
	DWD	61.45±2.53(7)	73.01±2.03(9)	77.32±0.93(9)	84.98±2.05(10)
GDA	SVM	79.09±2.04(8)	86.28±1.26(6)	88.98±2.56(5)	90.49±0.89(5)
	ML	76.52±3.21(3)	84.51±2.42(11)	86.43±1.67(4)	88.37±1.83(10)
	1NN	77.21±1.65(10)	87.24±1.58(10)	89.22±1.85(11)	90.53±1.62(13)
	DWD	80.08±1.82(9)	86.83±1.62(9)	89.25±0.94(10)	91.02±2.11(12)
Original features	SVM	80.27±0.94(204)	85.68±1.05(204)	86.02±0.87(204)	89.16±1.94(204)
	ML	---	---	---	---
	1NN	81.14±1.27(204)	82.41±1.38(204)	85.62±2.17(204)	87.19±0.84(204)
	DWD	81.13±1.11(204)	84.56±0.86(204)	87.03±0.68(204)	89.32±1.73(204)

References

[1] Wu, D., & Sun, D. (2013). Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review — Part II: Applications. Innovative Food Science & Emerging Technologies, vol. 19, pp. 15-28.

[2] Gao, J., Li, X., Zhu, F., & He, Y. (2013). Application of hyperspectral imaging technology to discriminate different geographical origins of Jatropha curcas L. seeds. Computers and Electronics in Agriculture, vol. 99, pp. 186-193.

[3] Bazi, Y., & Melgani, F. (2010). Gaussian Process Approach to Remote Sensing Image Classification. IEEE Transactions on Geoscience and Remote Sensing, vol. 48, no. 1, pp. 186–197.

[4] Zhong, Y., & Zhang L. (2012). An Adaptive Artificial Immune Network for Supervised Classification of Multi-/Hyperspectral Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing, vol. 50, no. 3, pp. 894–909.

[5] Landgrebe, D. A. (2003). Signal Theory Methods in Multispectral Remote Sensing, Hoboken, NJ: Wiley.

[6] Hughes, G. F. (1998). On the mean accuracy of statistical pattern recognition. IEEE Transactions on Information Theory, vol. IT-14, no. 1, pp. 55–63.

[7] Hall, P., Marron, J. S., & Neeman, A. (2005). Geometric representation of high dimension, low sample size data. Journal of the Royal Statistical Society: Series B, vol. 67, Part 3, pp. 427-444.

[8] Jackson, Q., & Landgrebe, D. A. (2001). An adaptive classifier design for high dimensional data analysis with a limited training data set. IEEE Transactions on Geoscience and Remote Sensing, vol. 39, no. 12, pp. 2664–2679.

[9] Camps-Valls, G., & Bruzzone, L. (2005). Kernel-based methods for hyperspectral image classification. IEEE Transactions on Geoscience And Remote Sensing, vol. 43, no. 6, pp. 1351–1362.

[10] Marron, J. S., Todd. M. J., & Ahn, J. (2007). Distance-Weighted Discrimination. Journal of the American Statistical Association, vol. 102, no. 480, pp. 1267-1271.

[11] Dhanjal, C., Gunn, S. R., & Shawe-Taylor, J. (2009). Efficient Sparse Kernel Feature Extraction Based on Partial Least Squares. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 8, pp. 1347-1361.

[12] Hosseini, S. A., & Ghassemian, H. (2015). Rational function approximation for feature reduction in hyperspectral data. Remote Sensing Letters, vol. 7, no. 2, pp. 101–110.

[13] Imani, M., & Ghassemian, H. (2015). High-Dimensional Image Data Feature Extraction by Double Discriminant Embedding. Pattern Analysis and Applications, DOI 10.1007/s10044-015-0513-z.

[14] Imani, M., & Ghassemian, H. (2016). Binary coding based feature extraction in remote sensing high dimensional data. Information Sciences, vol. 342, pp. 191-208.

[15] Jia, X., Kuo, B.C., & Crawford, M. (2013). Feature Mining for Hyperspectral Image Classification. Proceedings of the IEEE, vol. 101, no. 3, pp. 676-697.

[16] Maji, P., & Garai, P. (2013). Fuzzy–Rough Simultaneous Attribute Selection and Feature Extraction Algorithm. IEEE Transactions on Cybernetics, vol. 43, no. 4, pp. 1166-1177.

[17] Li, S., Qiu, J., Yang, X., Liu, H., Wan, D., & Zhu, Y. (2014). A novel approach to hyperspectral band selection based on spectral shape similarity analysis and fast branch and bound search. Engineering Applications of Artificial Intelligence, vol. 27, pp. 241-250.

[18] Esfandian, N., Razzazi, F., & Behrad, A. (2012). A clustering based feature selection method in spectro-temporal domain for speech recognition. Engineering Applications of Artificial Intelligence, vol. 25, no. 6, pp. 1194-1202.

[19] Dernoncourt, D., Hanczar, B., & Zucker, J.-D. (2014). Analysis of feature selection stability on high dimension and small sample data. Computational Statistics and Data Analysis, vol. 71, pp. 681–693.

[20] Liao, T. W. (2010). Feature extraction and selection from acoustic emission signals with an application in grinding wheel condition monitoring. Engineering Applications of Artificial Intelligence, vol. 23, no. 1, pp. 74-84.

[21] Zhang, Z., Tian, Z., Duan, X., & Fu, X. (2013). Adaptive kernel subspace method for speeding up feature extraction. Neurocomputing, vol. 113, pp. 58-66.

[22] Imani, M., & Ghassemian, H. (2015). Ridge regression-based feature extraction for hyperspectral data. International Journal of Remote Sensing, vol. 36, no. 6, pp. 1728–1742.

[23] Yu, G., & Kamarthi, S. V. (2010). A cluster-based wavelet feature extraction method and its application. Engineering Applications of Artificial Intelligence, vol. 23, no. 2, pp. 196-202.

[24] Li, B., Wang, C., & Huang, D. (2009). Supervised feature extraction based on orthogonal discriminant projection. Neurocomputing, vol. 73, pp. 191-196.

[25] Cariou, C., Chehdi, K., & Le Moan, S. (2011). BandClust: An Unsupervised Band Reduction Method for Hyperspectral Remote Sensing. IEEE Geoscience and Remote Sensing Letter, vol. 8, no. 3, pp. 565-569.

[26] Vicient, C., Sánchez, D., & Moreno, A. (2013). An automatic approach for ontology-based feature extraction from heterogeneous textualresources. Engineering Applications of Artificial Intelligence, vol. 26, no. 3, pp. 1092-1106.

[27] Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition, 2nd Ed, New York: Academic.

[28] Baudat, G., & Anouar, F. (2000). Generalized discriminant analysis using a kernel approach. Neural Computing, vol. 12, no. 10, pp. 2385–2404.

[29] Zortea, M., Haertel, V., & Clarke, R. (2007). Feature Extraction in Remote Sensing High-Dimensional Image Data. IEEE Geoscience and Remote Sensing Letter, vol. 4, no. 1, pp. 107 - 111.

[31] Yin, J., Gao, C., & Jia, X. (2012). Using Hurst and Lyapunov Exponent for Hyperspectral Image Feature Extraction. IEEE Geoscience and Remote Sensing Letter, vol. 9, no. 4, pp. 705 - 709.

[33] Yin, J., Gao, C., & Jia, X. (2013). Wavelet Packet Analysis and Gray Model for Feature Extraction of Hyperspectral Data. IEEE Geoscience and Remote Sensing Letter, vol. 10, no. 4, pp. 682 - 686.

[32] Kuo, B. C., & Landgrebe, D. A. (2004). Nonparametric weighted feature extraction for classification. IEEE Transactions on Geoscience and Remote Sensing, vol. 42, no. 5, pp. 1096-1105.

[33] Yang, J.-M., Yu, P.-T., Kuo, B.-C., & Huang, H.-Y. (2007). A Novel Non Parametric Weighted Feature Extraction Method for Classification of Hyperspectral Image with Limited Training Samples. IEEE Symposium on Geoscience and Remote Sensing, Barcelona, 2007.

[34] Huang, H.-Y., & Kuo, B.-C. (2010). Double Nearest Proportion Feature Extraction for Hyperspectral-Image Classification. IEEE Transactions on Geoscience and Remote Sensing, vol. 48, no. 11, pp. 4034–4046.

[35] Wen, J., Tian, Z., Liu, X., & Lin, W. (2013), Neighborhood Preserving Orthogonal PNMF Feature Extraction for Hyperspectral Image Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 6, no. 2, pp. 759-768.

[36] Lee, C., & Landgrebe, D. A. (1993), Feature Extraction Based on Decision Boundaries. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 4, pp. 388-400.

[37] Mendenhall, M. J., & Merényi, E. (2008), Relevance-Based Feature Extraction for Hyperspectral Images. IEEE Transactions on Neural Network, vol. 19, no. 4, pp. 658–672.

[38] Cai, D., He, X., & Han, J. (2007). Semi-supervised discriminant analysis. IEEE 11th International Conference on Computer Vision, Rio de Janeiro, 2007.

[39] Camps-Valls, G., & Vila, J. (2006). Composite Kernels for Hyper Spectral Image Classification. IEEE Geoscience and Remote Sensing Letter, vol. 3, no. 1, pp. 93–97.

[40] Archibald, R., & Fann, G. (2007). Feature selection and classification of hyperspectral images with support vector machines. IEEE Geoscience and Remote Sensing Letter, vol. 4, no. 4, pp. 674–677.

[41] Camps-Valls, G., Bandos Marsheva, T. V., & Zhou, D. (2007). Semi-Supervised Graph-Based Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing, vol. 45, no. 10, pp. 3044–3054.

[42] Theodoridis, S., & Koutroumbas, K. (2009). Pattern Recognition, 4th Ed. Elsevier Academic Press.

[43] Chang, C., & Linin, C. (2008). LIBSVM—A Library for Support Vector Machines, Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[44] Foody, G. M. (2004). Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. Photogrammetric Engineering & Remote Sensing, vol. 70, no. 5, pp. 627–633.

[45] Liao, W., Pižurica, A., Scheunders, P., Philips, W., & Pi, Y. (2013). Semisupervised Local Discriminant Analysis for Feature Extraction in Hyperspectral Images. IEEE Transactions on Geoscience and Remote Sensing, vol. 51, no. 1, pp. 184-198.

[46] Xia, J., Chanussot, J., Du, P., & He, X. (2014). (Semi-) Supervised Probabilistic Principal Component Analysis for Hyperspectral Remote Sensing Image Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing., vol. 7, no. 6, pp. 2225- 2237.

[47] Zhong, Y., Wu, Y., Xu, X., & Zhang, L. (2015). An Adaptive Subpixel Mapping Method Based on MAP Model and Class Determination Strategy for Hyperspectral Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 3, pp. 1411- 1426.

[48] Zhong, Y., Zhang, L., Huang, B., & Li, P. (2006). An Unsupervised Artificial Immune Classifier for Multi/Hyperspectral Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing, vol. 44, no. 2, pp. 420-431.

[49] Zhang, L., Zhong, Y., Huang, B., Gong, J., & Li, P. (2007). Dimensionality Reduction Based on Clonal Selection for Hyperspectral Imagery. IEEE Transactions on Geoscience and Remote Sensing, vol. 45, no. 12, pp. 4172- 4186.

[50] Imani, M. & Ghassemian, H. (2015). Feature reduction of hyperspectral images: Discriminant analysis and the first principal component**.** Journal of AI and Data Mining, vol. 3, no. 1, pp. 1-9.

[51] Hsu, P.-H. (2007). Feature extraction of hyperspectral images using wavelet and matching pursuit. ISPRS Journal of Photogrammetry & Remote Sensing, vol. 62, pp. 78–92.

[52] Li, F., Xu, L., Siva, P., Wong, A., & Clausi, D. A. (2015). Hyperspectral Image Classification With Limited Labeled Training Samples Using Enhanced Ensemble Learning and Conditional Random Fields. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, no. 6, pp. 2427- 2438.

[53] Imani, M. & Ghassemian, H. (2015). Overlap-based feature weighting for feature extraction of hyperspectral remote sensing imagery. Journal of AI and Data Mining, vol. 3, no. 2, pp. 181-190.

نشریه هوش مصنوعی و داده کاوی

استخراج ویژگی تصاویر ابرطیفی با استفاده از نمونه‌های نیمه برچسب‌دار مرزی و معیار ترکیبی

مریم ایمانی و حسن قاسمیان*

دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران.

ارسال ۲۰۱۶/۰۴/۲۷؛ پذیرش ۲۰۱۶/۰۹/۲۱

چکیده:

استخراج ویژگی یک مرحله پیش پردازش مهم جهت طبقه‌بندی تصاویر ابرطیفی است. روش تحلیل ممیز خطی (LDA) در تعداد نمونه‌های آموزشی کم خوب کار نمی‌کند. به علاوه، LDA کارایی ضعیفی برای داده غیر گوسی دارد. LDA با یک معیار کلی بهینه می‌شود. در نتیجه، آن برای استفاده در داده چند-مده منعطف نیست. در این مقاله، یک روش استخراج ویژگی جدید پیشنهاد شده است که از نمونه‌های نیمه برچسب‌دار مرزی برای حل مسئله مجموعه نمونه آموزشی کوچک استفاده می‌کند. روش پیشنهادی که استخراج ویژگی ترکیبی بر مبنای نمونه‌های نیمه برچسب‌دار مرزی (HFE-BSL) نامیده شده است، از یک معیار ترکیبی حاوی هر دو معیارهای کلی و محلی برای استخراج ویژگی استفاده می‌کند. در نتیجه، روش پیشنهادی پایدار و انعطاف پذیر است. نتایج آزمایش‌ها بر روی یک داده چند طیفی ساختگی و سه داده ابرطیفی واقعی، کارایی خوب HFE-BSL را در مقایسه با چند روش مهم و پرکابرد استخراج ویژگی نشان می‌دهد.

کلمات کلیدی: استخراج ویژگی، تصویر ابرطیفی، نمونه‌های مرزی، معیار ترکیبی، طبقه‌بندی.