Title: A Spectral-Spatial Multi-Criteria Active Learning Technique for Hyperspectral Image Classification

This paper appears in: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Date of Publication: 20 September 2017

Author(s): S. Patra, K Bhardwaj, L. Bruzzone

Volume: 10, Issue: 12

Page(s): 5213 - 5227

DOI: 10.1109/JSTARS.2017.2747600

# A Spectral-Spatial Multi-Criteria Active Learning Technique for Hyperspectral Image Classification

Swarnajyoti Patra, *Member, IEEE*, Kaushal Bhardwaj, and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—Hyperspectral image classification with limited labelled samples is a challenging task and still an open research issue. In this article a novel technique is presented to address such an issue by exploiting dimensionality reduction, spectral-spatial information and classification with active learning. The proposed technique is based on two phases. Considering the importance of dimensionality reduction and spatial information for the analysis of hyperspectral images, Phase I generates the patterns corresponding to each pixel of the image using both spectral and spatial information. To this end, first, principal components analysis is used to reduce the dimensionality of an hyperspectral image, then extended morphological profiles are exploited. The spectral-spatial information based patterns generated by extended morphological profiles are used as input to the Phase II. Phase II performs the classification task guided by an active learning technique. This technique is based on a novel query function that uses uncertainty, diversity and cluster assumption criteria by exploiting the properties of $k$-means clustering, $K$-nearest neighbors algorithm, support vector machines and genetic algorithms. Experiments on three benchmark hyperspectral data sets demonstrate that the proposed method outperforms five state-of-the-art active learning methods.**

***Index Terms*—Active learning, classification, genetic algorithms, $k$-means clustering, mathematical morphology, support vector machines, remote sensing.**

## I. Introduction

HYPERSPECTRAL images (HSIs) are characterized by hundreds of bands acquired in contiguous spectral ranges and narrow spectrum intervals. They represent a very rich information source for a precise characterization and recognition of objects on the ground. In the past decades researchers devoted great attention to the classification of hyperspectral images for numerous applications, like the detailed classification of forest areas, the analysis of inland water and coastal zones, the analysis of natural risks, etc [1]. Due to the existence of a large number of bands, classification of HSI requires a sufficiently large number of training (labelled) samples in order to mitigate curse of dimensionality (or Hughes phenomenon) [2]. However, in most of the hyperspectral applications, the numbers of available labelled samples is scarce and very costly to collect. To address such a problem, dimensionality reduction of HSIs is widely used in the literature [3]–[10]. Dimensionality reduction decreases the number of the HSI spectral channels with the help of feature selection (extraction) techniques that select (extract) only non-redundant informative features which preserve discriminative properties of the data.

Although dimensionality reduction mitigates the curse of dimensionality problem, the classification results still rely on the quality of the available labelled samples. Due to the usually complex statistical distributions of the patterns belonging to different classes, informative labelled samples (i.e., the non-redundant samples which distinguish among different classes) are essential to train the classifier. Two recent approaches to HSI classification using limited labelled samples are semisupervised learning and active learning. Semisupervised learning incorporates both the labelled and unlabelled data into the training phase of a classifier to obtain better decision boundaries [11]–[15]. In contrast, active learning (AL) is a paradigm to reduce the labeling effort and optimize the performance of a classifier by including only most informative patterns (which have highest training information for supervised learning) into the training set. AL techniques are usually based on iterative algorithms. At each iteration, one or multiple most informative unlabelled patterns are chosen for manual labeling and the classification model is retrained with the additional labelled samples. The step of training and the step of assigning labels are iterated alternately until a stable classification result is obtained, i.e., the classification accuracy does not increase further by increasing the number of training samples. Accordingly, the classifier is trained only with the most informative samples, thus reducing the labeling cost. In the literature many studies have shown that AL is a promising approach to classification of HSI with limited labelled samples [16], [17].

The fundamental component of AL is the design of a query function that should incorporate a set of criteria for selecting the most informative patterns to label from an unlabelled pool $U$. Depending on the number of samples to be selected at each iteration, two kinds of AL methods exist in the literature: 1) those that select the single most informative sample at each iteration, and 2) those that select a batch of informative samples at each iteration. To avoid retraining the classifier for each new labelled sample added to the training set, batch mode AL methods are preferred in the remote sensing community. AL has been widely studied in the pattern recognition literature [18]–[22]. In the recent years, several AL techniques have been proposed for classification of multispectral and hyperspectral remote sensing images [23]–[36]. Mitra *et al.* [23] presented an AL technique by adopting a one-against-all (OAA) architecture of binary support vector machine (SVM) classifiers. They select batch of uncertain samples, one from each binary SVM, by considering that closest to the discriminating hyperplane. In

S. Patra and K. Bhardwaj are with the Department of Computer Science and Engineering, Tezpur University, 784 028 Tezpur, India (e-mail:swpatra@tezu.ernet.in; kauscsp@tezu.ernet.in).

L. Bruzzone is with the Department of Information Engineering and Computer Science, University of Trento, I-38123 Trento, Italy (e-mail:lorenzo.bruzzone@unitn.it).

[24], an AL technique is presented that exploits the maximum-likelihood classifier and the Kullback-Leibler divergence. It selects the unlabelled sample that maximizes the information gain between the a posteriori probability distribution estimated from the current training set and the training set obtained by including also that sample. In [25], two batch mode active learning techniques are proposed for classification of remote sensing images. The first one extends the SVM margin sampling method by selecting the samples that are closest to the separating hyperplane and associated with different closest support vectors. The second method is based on a committee of classifiers. The samples that have maximum disagreement among the committee of learners are selected. In [26], Demir *et al.* investigated several SVM-based batch mode AL techniques for the classification of remote sensing images. In [27], a batch mode AL technique based on multiple uncertainty for SVM classifiers is presented. Few cluster assumption based AL techniques are presented in [28]–[30]. A cost-sensitive AL method for the classification of remote sensing images is presented in [31] and extended in [32]. This method includes in the query function also the cost associated with the accessibility of the unlabelled samples. An AL technique based on a Gaussian process classifier for hyperspectral image analysis is presented in [33]. All the above-mentioned AL methods only exploit spectral information. There are few techniques existing in the literature that exploit spectral and spatial information to achieve improved classification results [34]–[38].

As mentioned before feature selection (or extraction) plays an important role for HSI classification with limited labelled samples. Moreover, in practice, pixels are spatially related due to the homogeneous spatial distribution of land covers. It is highly probable that two adjacent pixels belong to the same class. Thus, information captured in neighboring locations may provide useful supplementary knowledge for analysis of a pixel. Therefore, spectral information with the support of spatial information can effectively reduce the uncertainty of class assignment and help to find the most informative samples.

In this work we propose a novel technique for the classification of HSI with limited labelled samples. The proposed technique is divided into two phases. Considering the importance of dimensionality reduction and spatial information for the analysis of HSIs, Phase I extracts the features corresponding to each pixel of HSI using both spectral and spatial information. To this end, first principal components analysis (PCA) is used to reduce the dimensionality of HSI; then, extended morphological profiles (EMP) are exploited. The spectral-spatial features based patterns (samples) generated by the EMP are used as input to the Phase II. Phase II performs the classification task with a small number of labelled samples. To this end, a multi-criteria batch mode AL technique is proposed by defining a novel query function that exploits the properties of the $k$-means clustering, the $K$-nearest neighbors algorithm, the SVM classifier, and genetic algorithms (GAs). The method first partitions the unlabelled pool $U$ generated by Phase I into a large number of clusters using the $k$-means clustering algorithms. Then by exploiting the properties of the $k$-means clustering and the $K$-nearest neighbors algorithms, for each $x \in U$ the density of the region in which the pattern $x$ belongs is computed. This density is used to incorporate the cluster assumption[1] criterion in the query function. The proposed technique also incorporates uncertainty and diversity criteria to select the informative samples at each iteration of AL. The uncertainty criterion is defined by exploiting an SVM classifier and the diversity criterion is defined by maximizing the nearest neighbor distances of the selected samples. In the proposed AL technique, at each iteration the SVM classifier is trained with the available labelled samples. After training, $m$ most uncertain samples are selected. Then, a batch of $h(h < m)$ informative samples from the selected $m$ samples are chosen for manual labeling by optimizing the uncertainty, diversity and cluster assumption criteria with GAs. To assess the effectiveness of the proposed method we compared it with five other batch mode AL techniques existing in the literature using three hyperspectral remote sensing data sets.

The rest of this paper is organized as follows. The proposed active learning technique is presented in Section II. Section III provides the description of the three hyperspectral remote sensing data sets used for experiments. Section IV presents the experimental results obtained on the considered data sets. Finally, Section V draws the conclusion of this work.

## II. Proposed Technique

In this paper we propose a technique for classification of HSIs with limited labelled samples. The proposed technique is divided into two phases. Phase I generates the patterns corresponding to each pixel of the HSIs by extracting spectral-spatial features. Phase II performs the classification task by exploiting a novel AL technique. Fig. 1 shows the block diagram of the proposed framework. The details steps of the proposed technique are given in next subsequent subsections.

### A. Phase I: spectral-spatial feature extraction

The classification of an HSI when a limited number of labelled samples is available is a challenging task due to the curse of dimensionality problem. Moreover, due to the existence of large number of redundant and irrelevant bands, the distributions of different classes in the original feature space are complex and do not follow the cluster assumption property, i.e. the interclass differences between classes are not significant. Thus, cluster assumption criterion may fail to play a significant role for identifying informative samples. Both problems can be solved by reducing the dimensionality of the HSI data by selecting (or extracting) only discriminative features. When a small number of discriminative features is considered, the class distributions might be much simpler and result in more significant interclass differences. Thus, finding the unlabelled informative samples in the reduced feature space is much easier than the original feature space. Moreover,

[1]The cluster assumption is equivalent to the low-density separation assumption which states that the decision boundary among classes should lie on a low-density region of the feature space. According to this assumption, one can say that two points in the feature space are likely to have the same class label if there is a path connecting them passing through high-density regions only [39].

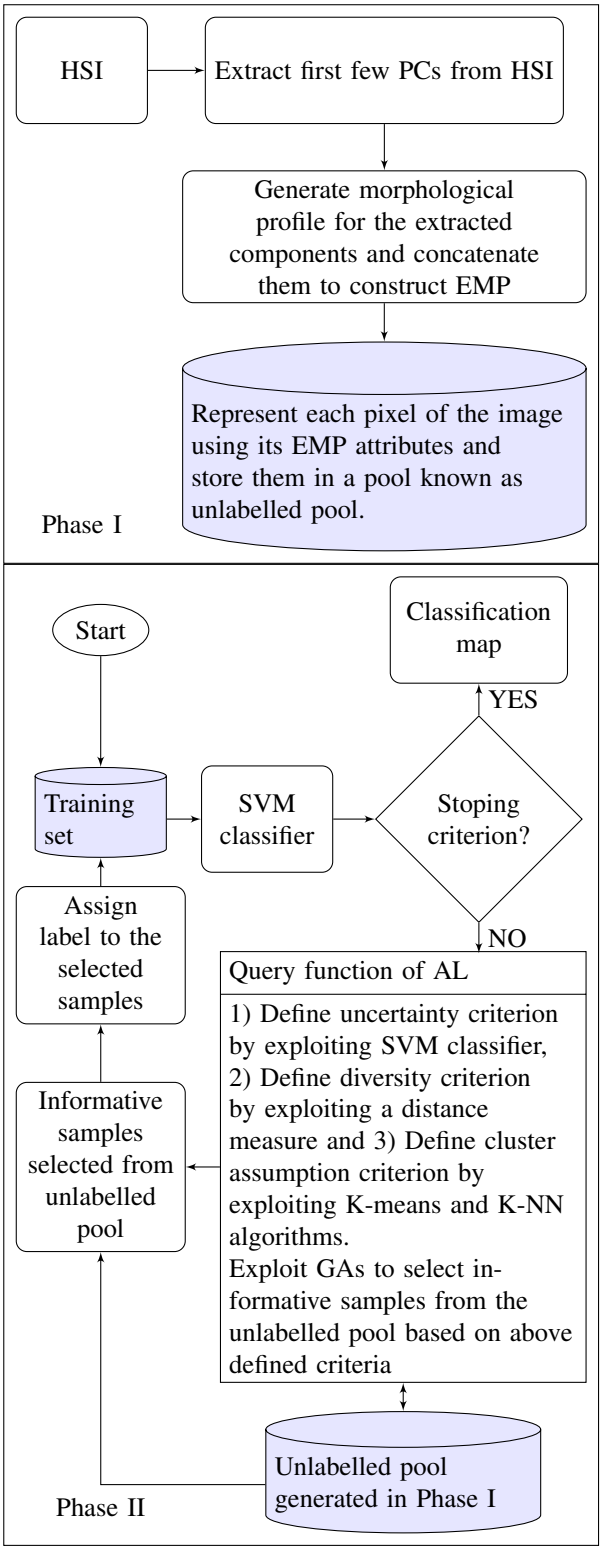


Fig. 1. Block diagram of the proposed framework.

a small number of informative labelled samples may be good enough to train a classifier. In the proposed technique we reduce the dimensionality of HSIs by extracting informative features with the help of principal component analysis.

*1) Principal component analysis:* PCA is an orthogonal transformation technique widely used in feature extraction and data compression [40]. It transforms a set of patterns in a $d$-dimensional original feature space into a new feature space having the same dimension where the transformed features are called principal components (PCs). The transformation is defined in such a way that the first PC has the largest possible variance of the patterns, and each succeeding component in turn has the highest variance possible under the constraint to be orthogonal to the preceding components. Thus, PCA orders the PCs according to the variance of the patterns. It is used to reduce dimensionality of the data by keeping first few PCs. In our work, the dimensionality of HSIs is reduced by keeping only the first $l$ PCs that retain more than 99% of information and discard the rest.

The spectral features extracted by PCA are not enough to distinguish classes in HSIs. In many HSIs pixels are spatially correlated due to the homogeneous spatial distribution of land covers. Information captured in neighboring pixels may provide useful supplementary knowledge for the analysis of a pixel. Therefore, spectral information with the support of spatial information can effectively reduce the uncertainty of class assignment and help the AL process to select more informative samples for labelling. In this work, EMP are used to incorporate spatial information into the extracted spectral features.

*2) Extended morphological profiles:* Mathematical morphology has been successfully applied to images [41]–[46]. Two fundamental morphological operators are dilation and erosion. Dilation $\delta_E(I)$ in a grey scale image $I$ replaces the pixel intensity with the maximum intensity value present in its neighborhood, which is defined by the structuring element $E$. The structuring element (SE) is a small structure which sets boundary of neighborhood for a pixel to be investigated. By duality erosion $\varepsilon_E(I)$ replaces the pixel by minimum intensity.

Two important morphological filters are opening and closing. Opening $\gamma_E(I)$ of an image $I$ by a structuring element $E$ is defined as the erosion of $I$ followed by the dilation with the symmetrical structuring element $E$.

$$\gamma_E(I) = \delta_E[\varepsilon_E(I)]$$

Closing $\phi_E(I)$ of an image $I$ by a structuring element $E$ is defined as the dilation of $I$ by $E$ followed by the erosion with symmetric SE.

$$\phi_E(I) = \varepsilon_E[\delta_E(I)]$$

When opening or closing is applied to an image, structures smaller than SE disappears. These filters may introduce false structures or modify existing structures which can be avoided by geodesic reconstruction. The composition of erosion and reconstruction by dilation is called opening by reconstruction or geodesic opening. The composition of dilation and reconstruction by erosion is called closing by reconstruction

or geodesic closing. On applying geodesic opening or closing we get a similar image with some preserved objects. Using varying size of SE we get multiple similar images (called granulometry) preserving shapes and size of all objects present in the image. A granulometry generated by geodesic opening using a SE of an increasing size is called opening profile (OP). Similarly, a granulometry generated by geodesic closing using a SE of an increasing size is called closing profile (CP). OP of image $I$ can be define as:

$$OP(I) = \{\gamma_R^{E_1}(I), \gamma_R^{E_2}(I), ..., \gamma_R^{E_t}(I)\}$$

where $t$ is the number of opening by reconstructions. $\gamma_R^{E_i}(I)$ is opening by reconstruction considering the size of structuring element $E_i$. It is defined as:

$$\gamma_R^{E_i}(I) = R_\delta^{E_i}[\varepsilon_{E_i}(I)]$$

where $R_\delta^{E_i}$ is reconstruction by dilation. Similarly, CP of image $I$ can be defined as:

$$CP(I) = \{\phi_R^{E_1}(I), \phi_R^{E_2}(I), ..., \phi_R^{E_t}(I)\}$$

where

$$\phi_R^{E_i}(I) = R_\varepsilon^{E_i}[\delta_{E_i}(I)]$$

.

The morphological profiles (MP) of an image $I$ is the concatenation of image $I$ with its opening profile and its closing profile i.e., $MP(I) = \{I, OP(I), CP(I)\}$. Thus, MP of image $I$ is a collection of $2t+1$ similar images with different spatial information. For hyperspectral image one can integrate spectral and spatial information by generating MP for all bands and use them together but this would increase the dimension exponentially. To mitigate this problem, one option is to reduce the dimension of the original HSI and then integrate the generated MP for each image in the reduced dimension. In the literature this is called extended morphological profiles.

In our work, as explained in subsection II-A, the dimensionality of HSI is reduced by PCA selecting first $l$ PCs. Then the EMP for a hyperspectral image $H$ are generated by concatenating the MP of $l$ different images generated by $l$ PCs.

$$EMP(H) = \{MP(PC_1), MP(PC_2), ..., MP(PC_l)\}$$

This results in $l(2t+1)$ images containing spectral-spatial information to represent the pixels of HSIs. Thus, using EMP patterns corresponding to the pixels of HSI are modeled with $l(2t+1)$ spectral-spatial features.

### *B. Phase II: proposed active learning technique for classification of HSIs*

To incorporate spectral-spatial information in the classification process, the feature vectors generated in Phase I are used as input to Phase II. In this phase, a novel batch mode AL technique is proposed for classification of HSI with limited labelled samples. In order to select the most informative samples to be labelled, the query function of our AL technique is designed based on uncertainty, diversity and cluster assumption criteria. The uncertainty criterion is defined by exploiting SVM classifier. The diversity criterion is defined by maximizing the nearest neighbor distances of the selected samples. The cluster assumption criterion is defined by using the properties of k-means clustering and nearest neighbor algorithms. Finally GAs are exploited to select batch of most informative samples by optimizing these criteria. The details of the proposed technique are given below.

*1) Uncertainty criterion:* In this work a one-against-all (OAA) SVM architecture, which involves $n$ binary SVMs (one for each information class), is adopted to define uncertainty criterion as well as to perform the classification task [47]. The uncertainty criterion aims at selecting the samples that have the lowest classification confidence among the unlabelled samples. To this end, at each iteration of AL, $n$ binary SVM classifiers are trained with the available labelled samples. After training, $n$ functional distances $f_i(x), i = 1, 2, ..., n$ are obtained, that correspond to the $n$ decision hyperplanes. Then, the classification confidence of each unlabelled sample $x \in U$ is associate with its uncertainty measure. The samples which have lower classification confidence are considered more uncertain. In the literature two alternative strategies are used for computing the classification confidence. The first strategy is based on the widely used marginal sampling (MS) technique, where the smallest distance among the $n$ decision hyperplanes is considered to compute the classification confidence of each unlabelled sample [23]. The second strategy, which is also used in our work, is based on the multiclass label uncertainty (MCLU) [26]. In MCLU, the difference between the first and second largest distance values to the hyperplanes is considered to compute the classification confidence $cc(x)$ of each unlabelled sample $x \in U$ as follows:

$$\begin{aligned} r_{max1} &= \arg \max_{i=1,...,n} \{f_i(x)\} \\ r_{max2} &= \arg \max_{\substack{j=1,...,n \\ j \neq r_{max1}}} \{f_j(x)\} \\ cc(x) &= f_{r_{max1}} - f_{r_{max2}} \end{aligned} \quad (1)$$

Thus, in the MCLU strategy, the classification confidence is assessed based on the two most likely classes to which the test pattern belongs. If the value of $cc(x)$ is high, the sample $x$ is assigned to the $r_{max1}$ class with high confidence. On the contrary, if $cc(x)$ is small, the sample $x$ is very close to the boundary between classes $r_{max1}$ and $r_{max2}$. Thus, its classification confidence for the $r_{max1}$ class will be low.

*2) Diversity criterion:* The samples selected using the uncertainty criterion may have high redundancy. The diversity criterion plays an important role to reduce this redundancy. It selects the samples from the already selected uncertain samples which are diverse from each other. The diversity criteria based on angle, closest support vector, clustering *etc.* are widely used in the AL literature [25], [26], [48]. In this work a simple criterion that maximizes the distance between sample and its nearest sample is used to select diverse samples. Let $x_1, x_2, ..., x_m$ be the $m$ most uncertain samples

selected from $U$ using the MCLU criterion defined above. Now the optimization of the following criterion is used to select $h(h < m)$ diverse samples from the selected $m$ samples:

$$\max \left\{ \sum_{i=1}^{h} \min_{\substack{i \neq j \\ j=1,\ldots,m}} \{d(x_i, x_j)\} \right\} \quad (2)$$

where $d(x_i, x_j)$ is the euclidian distance between the sample $x_i$ and $x_j$. The $h$ samples selected using (2) are diverse from each other since the criterion maximize the distance between each sample with its nearest sample.

*3) Cluster assumption criterion:* Cluster assumption property states that the decision boundary among classes should lie on a low-density region of the feature space. Thus, the patterns that belong to low density regions of the feature space are the most informative for a classifier. The density of a region to which a specific pattern belongs can be computed by taking the average distance from its K-nearest neighbor patterns. Such a way to compute the density for each unlabelled pattern is impractical and cumbersome. In this work we exploit the properties of $k$-means clustering to solve this problem.

Clustering is based on unsupervised learning for grouping a set of patterns in such a way that samples in the same group (called a cluster) are more similar to each other than to those in other groups. $k$-means clustering aims to partition the patterns into $k$ clusters in which each samples belongs to the cluster with the nearest mean, serving as a representative (prototype) of the cluster [49]. In our method, before iterative AL process is started, unlabelled patterns are partitioned into a large number of clusters and the prototype of each cluster is derived by using the $k$-means algorithm. Let $C_1, C_2, ..., C_k$ and $\mu_1, \mu_2, ..., \mu_k$ be the $k$ clusters and their corresponding representatives obtained by the $k$-means algorithm. Now the density of the region in which a cluster $C_i$ belongs can be computed as follows:

$$den(C_i) = \frac{1}{K} \sum_{x_i \in K-NN(\mu_i)} d(\mu_i, x_i) \quad (3)$$

where $K - NN(\mu_i)$ represents the $K$ neighbor patterns that are nearest to the cluster representative $\mu_i$. After finding the density of all clusters, the density of a region where a pattern $x_j$ belongs, denoted as $den(x_j)$, is computed as:

$$den(x_j) = den(C_i), \text{where } x_j \in C_i \quad (4)$$

According to the cluster assumption, the patterns having higher density values have higher probability to be in a low-density region in the feature space as compared to the patterns having lower density values. Thus, the density computed by (4) can be used to evaluate the cluster assumption property in the AL query.

*4) Selecting informative samples using GAs:* In this section a query strategy for AL based on the above-defined criteria is presented by exploiting GAs [50]. At each iteration of AL, first the $m$ samples from U that have the lowest classification confidence computed using (1) are selected. After that, the $h(h < m)$ most informative samples from the selected $m$ uncertain samples are chosen by optimizing uncertainty, diversity and cluster assumption criteria using GAs. The basic steps of GAs to select $h$ informative samples are described below.

**Chromosome representation:** Each chromosome is a sequence of binary numbers representing the $h$ samples. If $s$ bits are used to represent a sample, the length of a chromosome that represent $h$ samples will be $h \times s$ bits. The first $s$ bits of the chromosome represent the first sample, the next $s$ bits represent the second sample, and so on.

**Population initialization:** A collection of chromosomes is called population. The number of chromosomes belonging to a population defines the size of the population. A population is formed by generating a set of chromosomes. Each chromosome in the population is initialized randomly to represent $h$ samples.

**Fitness computation:** Design of an appropriate fitness function is the most important and challenging task of GAs, since the chromosomes of the population contain useful solutions by optimizing their fitness value. The fitness function $F(.)$ is also known as objective function. In this work the fitness function of the GA that compute the fitness values of the chromosomes is defined as follows:

$$F(x_1, x_2, ..., x_h) = \frac{1}{h} \sum_{i=1}^{h} cc(x_i) - \frac{1}{h} \left\{ \sum_{i=1}^{h} \min_{\substack{i \neq j \\ j=1,\ldots,m}} \{d(x_i, x_j)\} \right\} - \frac{1}{h} \sum_{i=1}^{h} den(x_i) + P \quad (5)$$

Here $h(h < m)$ informative samples are chosen from the $m$ uncertain samples (obtained by using the uncertainty criterion defined in (1)) by minimizing the objective function. The first, second and third terms of the above objective function compute the average classification confidence (using the uncertainty criterion defined in (1)), the average minimum neighbor distance (using the diversity criterion defined in (2)) and the average density (using the cluster assumption criterion defined in (4)) of the $h$ samples represented by a chromosome, respectively. If a sample appear multiple times in a chromosome, the parameter $P$ has a positive constant value as a penalty, otherwise it is zero. The smaller value of the first term and the larger values of second and third terms provide smaller values of the objective function. Thus minimizing the objective function defined in (5) a GA results in the selection of the most informative samples to be labelled for AL.

**Selection:** The selection process selects chromosomes from the mating pool directed by the survival of the fittest concept of natural genetic systems. The 'stochastic uniform' selection strategy has been adopted here.

**Crossover:** Crossover exchanges information between two parent chromosomes for generating two child chromosomes. Given a chromosome of length $h \times s$, a crossover point is randomly generated in the range $[1, s \times h - 1]$.

**Mutation:** Each chromosome undergoes mutation with a fixed probability. Given a chromosome in the population, a bit position (or gene) is mutated by simply flipping its value.

**Termination criterion:** The processes of fitness value computation for each chromosome in the population, selection, crossover, and mutation are executed for a maximum number of iterations or the number of iteration until the average fitness value of the population becomes stable.

After termination criterion is satisfied, the chromosome in the population that has the best fitness value is considered and the $h$ samples that belong to that chromosome are selected as informative samples for the AL. Algorithm 1 provides the details of the proposed AL technique.

**Algorithm 1** Proposed active learning technique

**Phase I**

1: Apply PCA to HSI and select first $l$ PCs.
2: Obtain MP of $l$ images generated by the $l$ PCs.
3: Generate EMP of the HSI by concatenating all the MPs obtained in the previous step.
4: Obtain the patterns (samples) associated with the pixels of HSI by using its EMP attributes.

**Phase II**

1: Apply $k$-means clustering algorithm to the samples generated by Phase I to obtain $k$ clusters and their representatives.
2: Compute the density of each cluster using (3) and then for each $x \in U$ compute the local density of the region of the feature space in the neighborhood of $x$ by using (4).
3: **repeat**
4: Train binary SVMs in the OAA architecture with the available training samples and compute the classification confidence of each unlabelled sample $x \in U$ by using (1).
5: Select the $m(m < k)$ samples from $U$ that have the lowest classification confidence.
6: Exploit GAs to select a batch of $h(h < m)$ informative samples from $m$ by minimizing the objective function defined in (5).
7: Assign labels to the $h$ selected samples and include them into the training sat.
8: **until** the stop criterion is satisfied.

## III. Description of Data Sets

In order to assess the effectiveness of the proposed technique, three hyperspectral data sets were used in the experiment. The first data set[2] shown in Fig. 2 is a hyperspectral image acquired on the Kennedy Space Center (KSC), Merritt Island, Florida, USA, on March 23, 1996. This image consists of 512 x 614 pixels and 224 bands with a spatial resolution of 18 m. The number of bands is initially reduced to 176 by removing water absorption and low signal-to-noise bands. The labelled data were collected using land-cover maps derived from color infrared photography provided by KSC and Landsat thematic mapper imagery. The class name and corresponding numbers of ground truth observations used in the experiments are listed in Table I.

The second data set[2] is a hyperspectral image acquired by the ROSIS-03 (Reflective Optics System Imaging Spectrometer) optical sensor over the urban area of the University of Pavia, Italy. The flight was operated by the Deutsches Zentrum für Luft-und Raumfahrt (DLR, the German Aerospace Center)

[2]Available online: http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

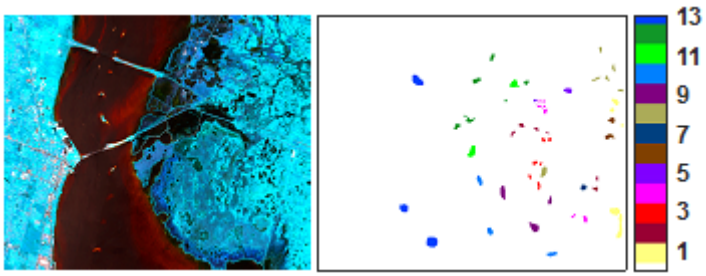


Fig. 2. Hyperspectral KSC image and its reference map.

TABLE I
KSC Data Set: Class Numbers, Class Names and Corresponding Numbers of Ground Truth Observations

| Class no. | Class name | No of labelled samples |
|---|---|---|
| 1 | Scrub | 761 |
| 2 | Willow swamp | 243 |
| 3 | Cabbage palm hammock | 256 |
| 4 | Cabbage palm/Oak hammock | 252 |
| 5 | Slash pine | 161 |
| 6 | Oak/Broadleaaf hammock | 229 |
| 7 | Hardwood swamp | 105 |
| 8 | Graminoid marsh | 431 |
| 9 | Spartina marsh | 520 |
| 10 | Cattaial marsh | 404 |
| 11 | Salt marsh | 419 |
| 12 | Mud flats | 503 |
| 13 | Water | 927 |
| **Total** | | **5211** |

in the framework of the HySens project, managed and funded by the European Union [51]. The size of the image in pixels is 610 x 340, with very high spatial resolution of 1.3 m/pixel. The number of bands of the ROSIS-03 sensor is 115 with a spectral coverage ranging from 430 to 860 nm. Some channels (twelve) have been removed due to noise. The remaining 103 spectral bands are processed. Fig. 3 shows a false color composite of the image. The class name and corresponding numbers of ground truth observations used in the experiments are listed in Table II.

TABLE II
Pavia University Data Set: Class Numbers, Class Names and Corresponding Numbers of Ground Truth Observations

| Class no. | Class name | No of labelled samples |
|---|---|---|
| 1 | Asphalt | 6631 |
| 2 | Meadows | 18649 |
| 3 | Gravel | 2099 |
| 4 | Trees | 3064 |
| 5 | Metal Sheets | 1345 |
| 6 | Bare Soil | 5029 |
| 7 | Bitumen | 1330 |
| 8 | Self-Blocking Bricks | 3682 |
| 9 | Shadows | 947 |
| **Total** | | **42776** |

The third data set[3] is an hyperspectral image acquired by the AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) sensor over the agricultural land of Indian Pines, Indiana, in

[3]Available online: http://engineering.purdue.edu/ biehl/MultiSpec

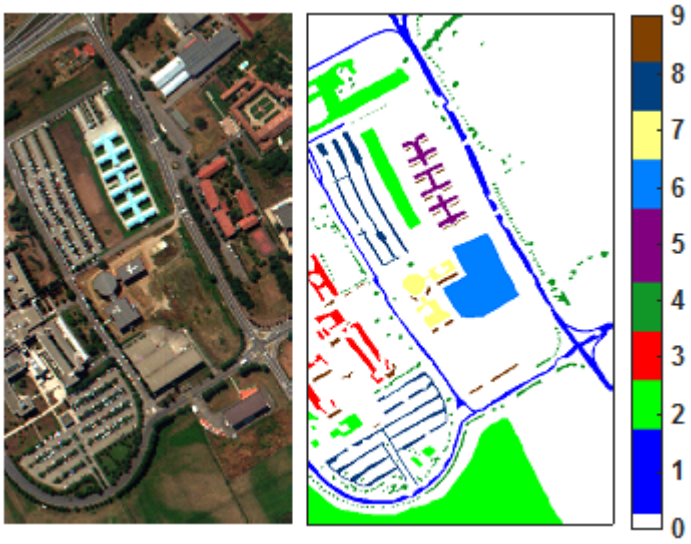


Fig. 3. Hyperspectral Pavia University image and its reference map.

TABLE III
INDIAN PINES DATA SET: CLASS NUMBERS, CLASS NAMES AND CORRESPONDING NUMBERS OF GROUND TRUTH OBSERVATIONS

| Class no. | Class name | No of labelled samples |
|---|---|---|
| 1 | Alfalfa | 46 |
| 2 | Corn-notill | 1428 |
| 3 | Corn-min | 830 |
| 4 | Corn | 237 |
| 5 | Grass/Pasture | 483 |
| 6 | Grass/Trees | 730 |
| 7 | Grass/Pasture-mowed | 28 |
| 8 | Way-windrowed | 478 |
| 9 | Oats | 20 |
| 10 | Soybeans-notill | 972 |
| 11 | Soybeans-min | 2455 |
| 12 | Soybean-clean | 593 |
| 13 | Wheat | 205 |
| 14 | Woods | 1265 |
| 15 | Bldg-Grass-Tree-Drives | 386 |
| 16 | Stone-steel towers | 93 |
| **Total** | | **10249** |

the early growing season of 1992. These data were acquired in the spectral range 400-2500 nm with spectral resolution of about 10 nm. The image consists of 145 x 145 pixels and 220 spectral bands with a spatial resolution of 20 m. Twenty water absorption and fifteen noisy bands were removed and the remaining 185 bands were included as candidate features. This image is used as it is a well known benchmark in the hyperspectral community. Fig. 4 shows a false color composition of the AVIRIS Indian Pines scene. The class name and corresponding numbers of ground truth observations used in the experiments are listed in Table III.

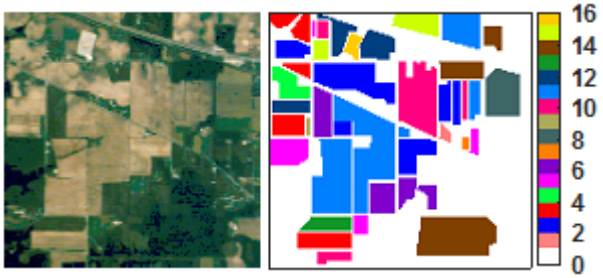


Fig. 4. Hyperspectral Indian Pines image and its reference map.

## IV. EXPERIMENTAL RESULTS

### A. Design of experiments

In order to show the potential of the proposed technique the three hyperspectral data sets described in Section III are used for experiments. Moreover, to assess the effectiveness of the proposed method, it is compared with four batch mode state-of-the-art AL methods existing in the literature: i) the entropy query-by-bagging (EQB) [25]; ii) the marginal sampling with angle based diversity (MS-ABD) [48]; iii) the cluster assumption with histogram thresholding (CAHT) [27]; and iv) the multiclass label uncertainty with enhanced cluster based diversity (MCLU-ECBD) [26]. The MS-ABD, the CAHT, and the MCLU-ECBD first select $m(m > h)$ most uncertain samples from $U$ by exploiting MS, CA and MCLU criteria, respectively. Then, by adopting different diversity criteria (the MS-ABD uses angle based diversity criterion, while the CAHT and the MCLU-ECBD use the kernel k-means clustering based diversity criterion) batches of $h(h >= 1)$ informative samples from the selected $m$ samples are chosen for labeling at each iteration of AL. In our experiments the value of $m$ is fixed to $3h$ for a fair comparison among the different techniques. The EQB technique directly selects the $h$ most uncertain samples according to the maximum disagreement between a committee of classifiers. The committee is obtained by bagging. Note that all the above-mentioned AL methods consider only spectral features as input. The proposed technique generated spectral-spatial features which are used as input to the AL process. In order to show the potential of the features generated by the proposed technique, the spectral-spatial features generated by our technique are also used as input to the above mentioned AL methods, referring to them as: i) SP-EQB; ii) SP-MS-ABD; iii) SP-CAHT; and iv) SP-MCLU-ECBD. Furthermore, to validate the effectiveness of the proposed technique, it also compared with an existing spectral-spatial information based state-of-the-art AL technique referred as MPM-LBP-BT technique [35]. The MPM-LBP-BT AL technique exploits spectral and spatial information by exploiting maximum *a posteriori* marginal (MPM) solution and loopy belief propagation. Then a breaking ties (BT) uncertainty criterion is used for query selection.

As explained in Section II the proposed technique reduces the dimensionality of the hyperspectral data by using PCA. In this experiment, the dimensionality of all the considered data sets are reduced by fixing the value of $l$ to 10 (i.e., only the first 10 PCs are considered and the remaining ones are discarded). To incorporate spatial information in the reduced dimension, an EMP with two opening and two closing leading to a stack of 50 features (5 for each PC) is computed to generate the patterns associated to the pixels of the HSI by considering disk-shape SE of radius 5 and 10. Thus, each pixel of the

hyperspectral image that is used as an input to our active learning is represented with 50 features containing spectral as well as spatial information.

To compute the density of the patterns in a specific region of the feature space, the proposed technique first partitions the feature space into a large number of clusters by using $k$-means clustering. Then the density of each cluster is computed by the $K$-nearest neighbors algorithm. In the experiments for all the data sets, the values of $k$ for $k$-means and $K$ for $K$-nearest neighbors algorithms are set to 500 and 10, respectively. The proposed technique also exploits GAs to select most informative samples. In our experiments for all the data sets the population size of GAs is taken as 20. Stochastic selection strategy is used to select fittest chromosomes from the mating pool. The crossover and mutation probabilities are set to 0.8 and 0.01 respectively.

All the active learning algorithms presented in this paper have been implemented in Matlab (R2015a). OAA SVM with radial basis function (RBF) kernels has been implemented by using the LIBSVM library [52]. The SVM parameters $\{\sigma, C\}$ (the spread of the RBF kernel and the regularization parameter) for all the data sets were derived by applying a grid search according to a five-fold cross-validation technique. The cross-validation procedure aimed at selecting the initial parameter values for the SVM. For simplicity, these values were not changed during the active learning iterations.

### B. Results: KSC data set

The first experiment is carried out to compare the performance of the proposed technique with the literature methods using the KSC data set. For this experiment, a total of $T = 5211$ labelled samples (see Table I) were considered as a test set TS. First, only 39 samples (three samples from each class) were randomly selected from $T$ as initial training set $L$, and the remaining 5172 were stored in the unlabelled pool $U$. At each iteration of AL 20 samples were selected from $U$ for labeling and the process was iterated 19 times resulting in 419 samples in the training set $L$. To reduce the random effect on the results, the active learning process was repeated for 10 trials with different initial labelled samples.

Fig.5 shows the average overall classification accuracies provided by the different methods versus the number of labelled samples included into the training set for the KSC data set. From this figure one can see that the EQB, the MS-ABD, the CAHT, and the MCLU-ECBD methods produced significantly higher classification accuracy when they use the spectral-spatial information based patterns included in the proposed technique as input instead of the patterns generated by considering only spectral bands. If the input patterns are generated by EMPs, these AL methods increase their accuracy of about 5%. This shows the importance of the spatial information for achieving better classification results. It is worth noting that as at the initial stage of AL the SVM decision hyperplane is far from the optimal hyperplane, the cluster assumption criterion of the proposed technique does not play significant role to select informative samples. As a result, at the initial iterations the proposed technique did not provide better

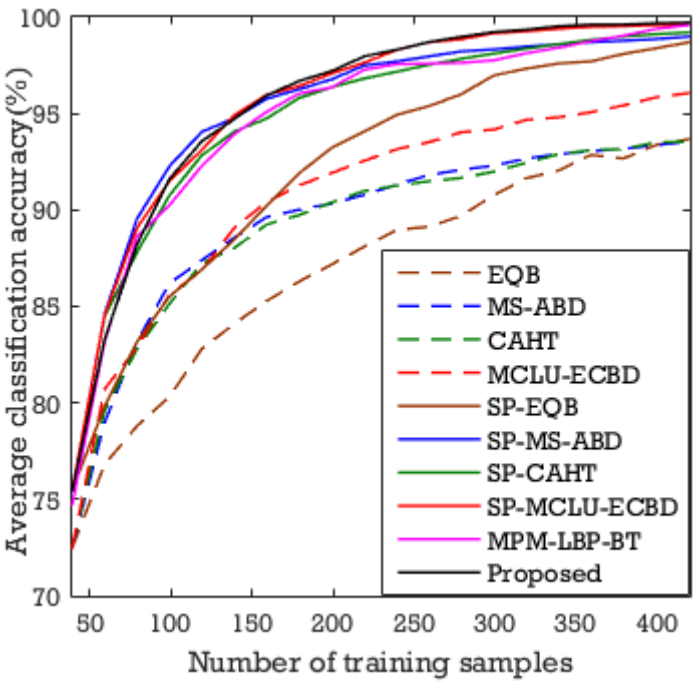


Fig. 5. Average classification accuracy over ten runs versus the number of training samples provided by the different methods (KSC data set).

results than the SP-MS-ABD technique. Nonetheless, after few iterations, the proposed technique outperformed all the existing AL techniques. Moreover, from the figure one can also see that the proposed technique always produced better results than the existing spectral-spatial information based state-of-the-art MPM-LBP-BT technique.

TABLE IV
AVERAGE OVERALL CLASSIFICATION ACCURACY ($\overline{OA}$), ITS STANDARD DEVIATION ($s$) AND KAPPA ACCURACY OBTAINED ON TEN RUNS FOR DIFFERENT TRAINING DATA SIZES (KSC DATA SET)

| **Methods** | $\lvert L\rvert = 239$ | | | $\lvert L\rvert = 339$ | | | $\lvert L\rvert = 419$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\overline{OA}$ | $s$ | kappa | $\overline{OA}$ | $s$ | kappa | $\overline{OA}$ | $s$ | kappa |
| SP-EQB | 94.93 | 1.15 | .943 | 97.59 | 0.63 | .973 | 98.59 | 0.45 | .985 |
| SP-MS-ABD | 97.70 | 0.42 | .974 | 98.59 | 0.45 | .984 | 98.98 | 0.42 | .989 |
| SP-CAHT | 97.16 | 0.39 | .968 | 98.59 | 0.27 | .984 | 99.20 | 0.13 | .991 |
| SP-MCLU-ECBD | 98.29 | 0.40 | .981 | 99.39 | 0.20 | .993 | 99.63 | 0.12 | .996 |
| MPM-LBP-BT | 97.55 | 1.18 | .973 | 98.41 | 1.25 | .982 | 99.58 | 0.26 | .995 |
| Proposed | **98.30** | **0.19** | **.981** | **99.53** | **0.10** | **.995** | **99.71** | **0.03** | **.997** |

To assess the effectiveness of the proposed AL method, Table IV shows the average overall classification accuracy (%), its standard deviation and the average kappa accuracies obtained by different AL techniques on ten runs with different numbers of training samples. From this table one can see that the proposed AL technique results in better classification accuracy than the other existing AL techniques. In particular, it is observed that the standard deviation of the proposed approach is always smaller than those of the other techniques. For example, considering 419 labelled samples, the proposed technique resulted in an overall accuracy of 99.71% with a standard deviation of 0.03. Whereas, among the literature methods, the highest overall accuracy produced by the SP-MCLU-ECBD technique was 99.63% with standard deviation

of 0.12. This confirms the better stability of the proposed method versus the choice of the initial training samples. It is worth noting that, the better results provided by the proposed technique are due to its capability to select the informative samples not only considering uncertainty and diversity criteria but also using the cluster assumption criterion. Table V shows the average class-wise accuracies (%) obtained by different AL techniques after completing 19 iterations (i.e., 419 samples in the training set $L$). From the table, one can see that for most of the classes the classification accuracies obtained by the proposed technique is either better or very closed to the best accuracy obtained by the literature methods. This shows that the integration of dimensionality reduction, spectral-spatial feature generation and the new query function of the AL method makes the proposed technique more robust not only to achieve higher classification accuracy but also to the quality of the initial training samples. For qualitative analysis Fig. 6 shows the classification maps obtained by the different AL techniques.

### C. Results: Pavia University data set

In order to assess the effectiveness of the proposed technique, the second experiment is carried out considering the Pavia University data set. For this experiment, $T = 42776$ labelled samples (see Table II) were considered as a test set TS. First only 27 samples (three samples from each class) were randomly selected from $T$ as training set L, and the remaining 42749 were stored in the unlabelled pool $U$. At each iteration of AL 20 samples were selected from $U$ for labeling and the process was iterated 19 times resulting in 407 samples in the final training set $L$. To reduce the random effect on the results, also in this one the active learning process was repeated for 10 runs with different initial labelled samples.

Fig.7 shows the average overall classification accuracies provided by the different methods versus the number of samples included into the training set. Similarly to the KSC data set, from this figure one can see that the classification results of the EQB, the MS-ABD, the CAHT, and the MCLU-ECBD significantly improved when considering as input the spectral-spatial information based patterns generated by the EMP included in the proposed technique. The increase in classification accuracy is of at least 7%. This again shows the importance of the spatial information for achieving better classification results. Furthermore, from the figure one can see that for the Pavia University data set, among the six spectral-spatial AL techniques, the MPM-LBP-BT technique provides worst classification results.

Table VI shows the average overall classification accuracy (%), its standard deviation and the average kappa accuracies obtained by different AL techniques on ten runs with different number of training samples. From this table one can see that the proposed AL technique produces better classification accuracy than the other existing AL techniques. For example, considering 407 labelled samples, the proposed technique resulted in an overall accuracy of 99.66% with standard deviation 0.04. Whereas, among the literature methods, the highest overall accuracy produced by the SP-MCLU-ECBD technique is 99.53% with standard deviation 0.19. The smaller standard deviation confirms the better stability of the proposed method versus the choice of the initial training samples. Table VII shows the average class-wise accuracies (%) obtained by different AL techniques after completing 19 iterations (i.e., 407 samples in the training set $L$). From the table, one can see that the class-wise average classification accuracies obtained by the proposed method are either better or comparable to the best results obtained by the literature methods. This shows the effectiveness of the proposed technique. Fig. 8 shows the classification maps obtained by the different AL techniques for visual analysis.

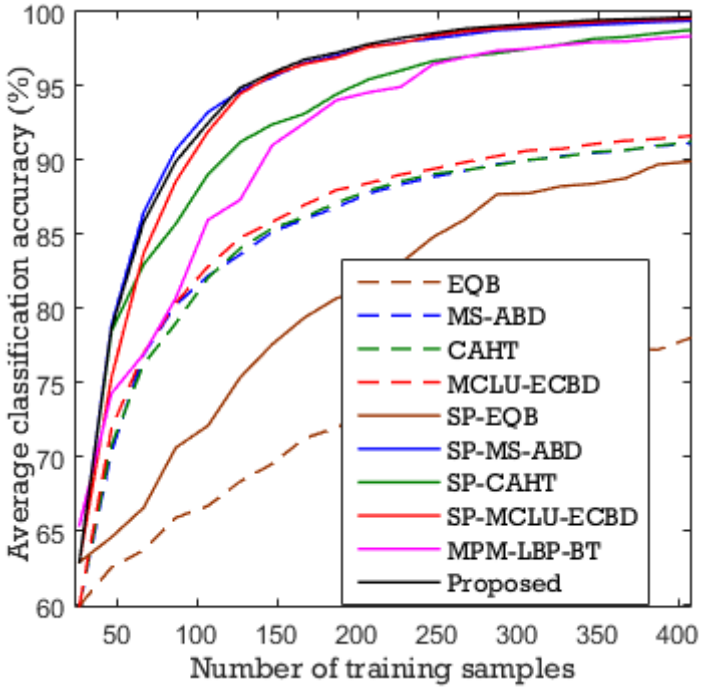


Fig. 7. Average classification accuracy over ten runs versus the number of training samples provided by the different methods (Pavia University data set).

TABLE VI
AVERAGE OVERALL CLASSIFICATION ACCURACY ($\overline{OA}$), ITS STANDARD DEVIATION ($s$) AND KAPPA ACCURACY OBTAINED ON TEN RUNS FOR DIFFERENT TRAINING DATA SIZES (PAVIA UNIVERSITY DATA SET)

| **Methods** | $\lvert L\rvert = 227$ | | | $\lvert L\rvert = 327$ | | | $\lvert L\rvert = 407$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\overline{OA}$ | $s$ | kappa | $\overline{OA}$ | $s$ | kappa | $\overline{OA}$ | $s$ | kappa |
| SP-EQB | 83.07 | 4.64 | .785 | 88.25 | 4.95 | .849 | 89.92 | 3.65 | .870 |
| SP-MS-ABD | 97.97 | **0.39** | .973 | 99.04 | 0.25 | .987 | 99.40 | 0.10 | .992 |
| SP-CAHT | 96.05 | 0.99 | .948 | 97.82 | 0.43 | .971 | 98.79 | 0.32 | .984 |
| SP-MCLU-ECBD | 97.91 | 0.51 | .972 | 99.17 | 0.38 | .989 | 99.53 | 0.19 | .994 |
| MPM-LBP-BT | 94.94 | 2.15 | 0.931 | 97.77 | 0.51 | 0.970 | 98.36 | 0.74 | 0.978 |
| Proposed | **98.24** | 0.46 | **.977** | **99.32** | **0.14** | **.991** | **99.66** | **0.04** | **.995** |

### D. Results: Indian Pines data set

In order to assess the effectiveness of the proposed technique, the third experiment is carried out considering the Indian Pines data set. A total of $T = 10249$ labelled samples (see Table III) are considered as a test set TS. For this experiment, first only 48 samples (three samples from each class) are randomly selected from $T$ as training set L, and the

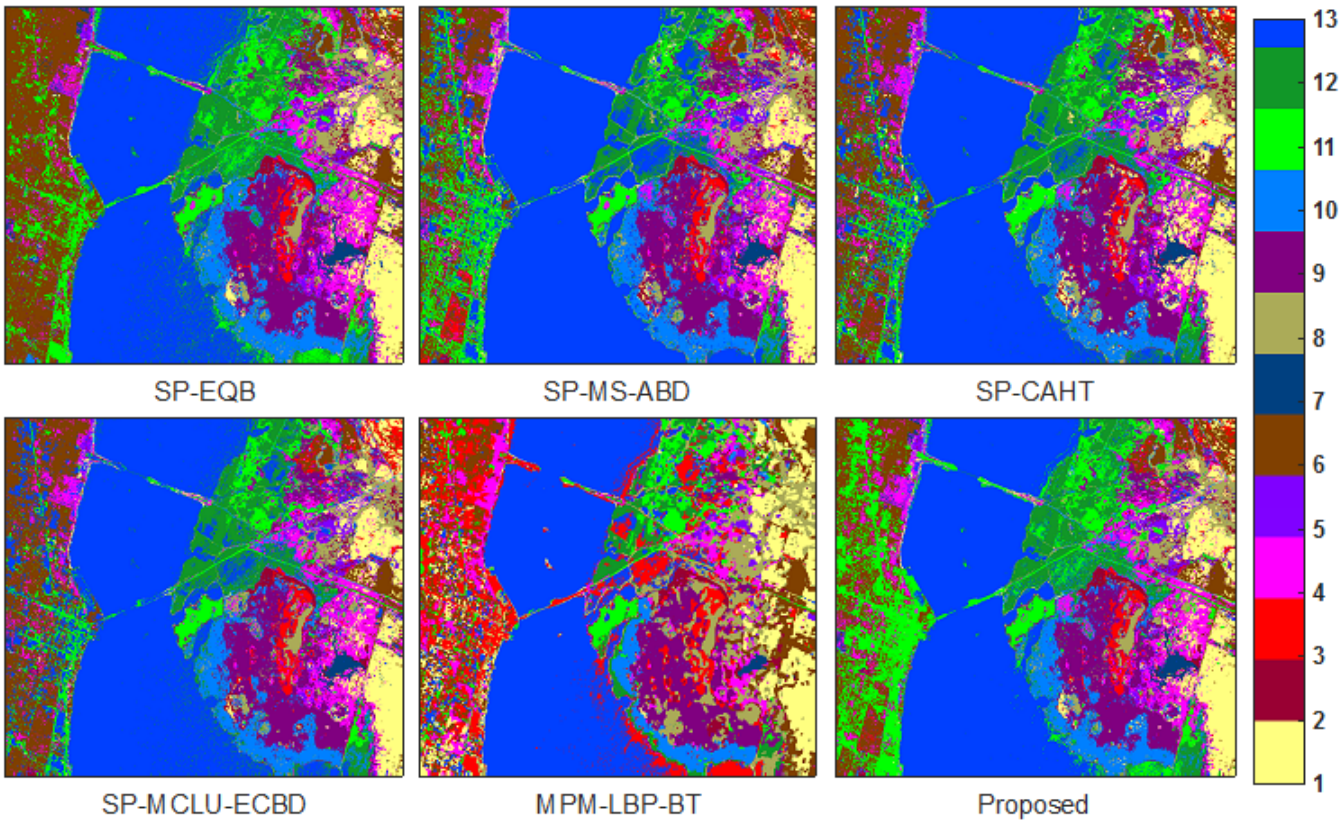


Fig. 6. Classification maps provided by different approaches with 419 labelled samples on the KSC data set.

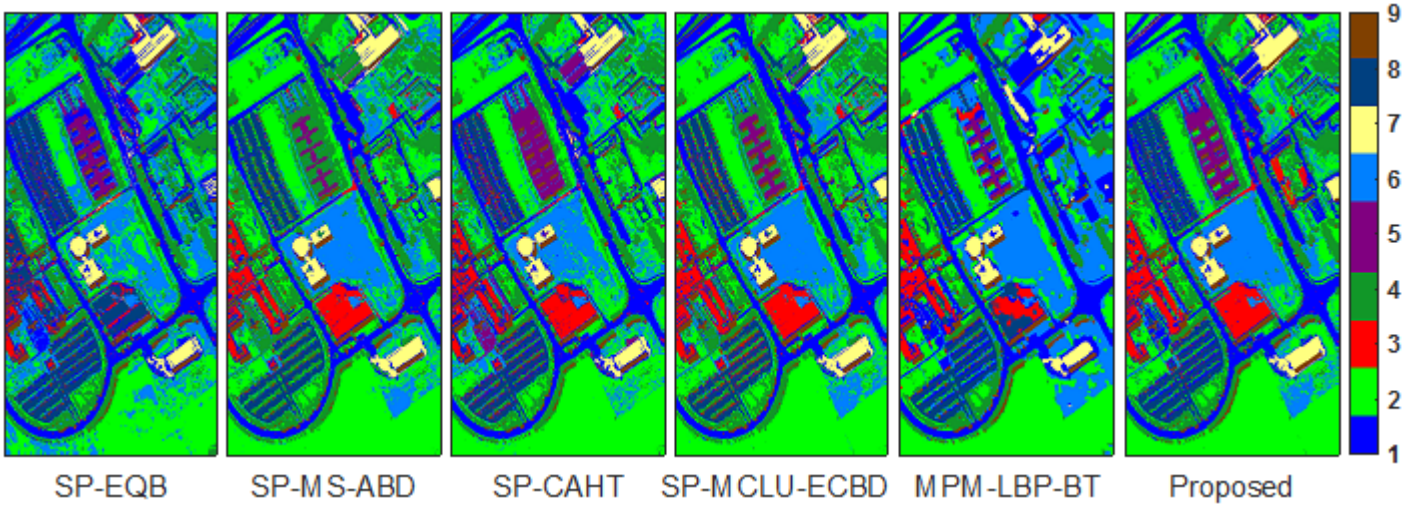


Fig. 8. Classification maps provided by different approaches with 419 labelled samples on the Pavia University data set.

TABLE V
CLASS WISE AVERAGE CLASSIFICATION ACCURACIES (%) OBTAINED ON TEN RUNS (KSC DATA SET).

| Methods | SP EQB | SP-MS ABD | SP CAHT | SP-MCLU ECBD | MPM LBP-BT | Proposed |
|---|---|---|---|---|---|---|
| $\|L\|$ | | | | **419** | | |
| **Scrub** | 99.98 | 99.96 | 99.92 | **100** | **100** | **100** |
| **Willow swamp** | 93.42 | 99.30 | 97.12 | 99.75 | **100** | **99.88** |
| **Cabbage palm hammock** | 98.83 | 99.53 | 99.22 | 99.96 | **100** | 99.92 |
| **Cabbage palm/Oak hammock** | 95.52 | 98.77 | 97.78 | **99.16** | 95.51 | 99.09 |
| **Slash pine** | 96.40 | 95.22 | 94.41 | 94.84 | **99.00** | 95.16 |
| **Oak/Broadleaaf hammock** | 99.61 | 99.96 | 99.13 | **100** | **100** | **100** |
| **Hardwood swamp** | 96.38 | 98.19 | 95.33 | 98.29 | **100** | 99.15 |
| **Graminoid marsh** | 99.84 | 99.95 | 99.68 | **100** | **100** | **100** |
| **Spartina marsh** | 99.94 | 99.90 | 99.85 | **99.98** | 99.70 | 99.96 |
| **Cattaial marsh** | 98.07 | 95.84 | 98.76 | 98.94 | 99.47 | **99.36** |
| **Salt marsh** | 99.64 | 99.52 | 99.88 | **99.93** | **100** | 99.88 |
| **Mud flats** | 98.21 | 96.98 | 99.62 | 99.64 | 98.51 | **99.92** |
| **Water** | 99.18 | 99.75 | 99.87 | **100** | **100** | **100** |
| $\overline{OA}$ | 99.59 | 98.98 | 99.20 | 99.63 | 99.58 | **99.71** |

TABLE VII
CLASS ACCURACIES (%), AVERAGE OVERALL CLASSIFICATION ACCURACY ($\overline{OA}$) AND ITS STANDARD DEVIATION ($std$), AND AVERAGE KAPPA ($kappa$) ACCURACY OBTAINED ON TEN RUNS (PAVIA UNIVERSITY DATA SET).

| Methods | SP EQB | SP-MS ABD | SP CAHT | SP-MCLU ECBD | MPM LBP-BT | Proposed |
|---|---|---|---|---|---|---|
| $\|L\|$ | | | | **407** | | |
| **Asphalt** | 99.21 | 99.19 | 99.07 | **99.56** | 99.30 | 99.51 |
| **Meadows** | 83.69 | 99.81 | 99.12 | 99.82 | 99.66 | **99.86** |
| **Gravel** | 84.87 | 98.17 | 96.66 | 98.17 | 89.87 | **98.82** |
| **Trees** | 99.04 | 99.32 | 98.61 | 99.38 | 98.67 | **99.56** |
| **Metal Sheets** | 96.73 | 99.85 | 99.93 | **99.96** | 92.43 | 99.94 |
| **Soil** | 86.70 | 98.78 | 97.92 | 99.32 | 99.53 | **99.68** |
| **Bitumen** | 96.59 | **99.26** | 97.89 | 99.01 | 94.95 | 99.17 |
| **Bricks** | 97.16 | 99.11 | 98.76 | 99.19 | 96.05 | **99.24** |
| **Shadows** | 99.07 | 99.88 | **99.90** | **99.90** | 99.68 | **99.90** |
| $\overline{OA}$ | 89.92 | 99.40 | 98.79 | 99.53 | 98.36 | **99.66** |

remaining 10201 are stored in the unlabelled pool $U$. At each iteration of AL, 20 samples are selected from $U$ for labeling and the process is iterated 45 times resulting in 948 samples in the training set $L$. Also in this case, the active learning process is repeated for 10 runs with different initial labelled samples.

Fig.9 shows the average overall classification accuracies provided by the different methods versus the number of samples included into the training set for Indian Pines data set. Also on this data set the classification accuracies of the EQB, the MS-ABD, the CAHT, and the MCLU-ECBD significantly improved (at least of 8%) when considering as input the spectral-spatial information based patterns generated by EMP. This again shows the effectiveness of spectral-spatial features generated by Phase I of the proposed technique. From the figure one can also see that at the initial iterations of the AL process the proposed technique provided better results than the existing MPM-LBP-BT technique.

Table VIII shows the average overall classification accuracy (%), its standard deviation and the average kappa accuracies obtained by different AL techniques with different number of labelled samples. From this table one can see that the proposed AL method produces second highest classification accuracy with lower standard deviations among the considered AL techniques. Although, for Indian Pines data set the MPM-LBP-BT technique resulted in the highest accuracy, it produced worst results for the KSC and the Pavia University data sets. Table IX shows the average class-wise accuracies (%) obtained by different AL techniques after completing 45 iterations (i.e.,

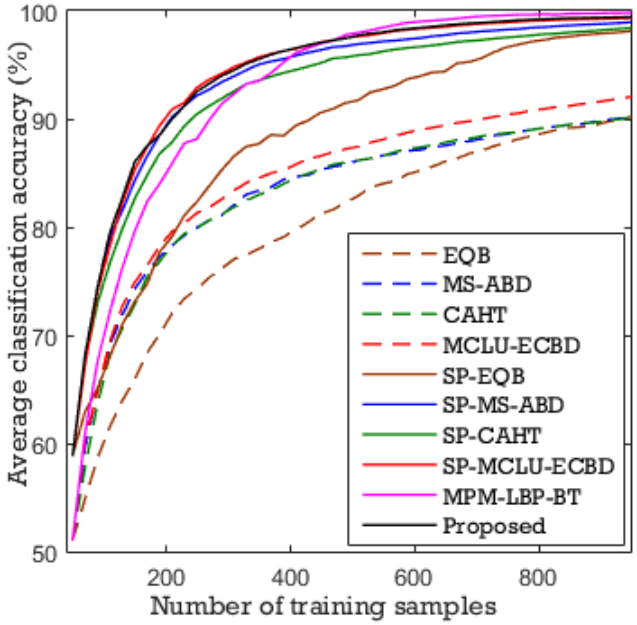


Fig. 9. Average classification accuracy over ten runs versus the number of training samples provided by the different methods (Indian Pines data set).

TABLE VIII
AVERAGE OVERALL CLASSIFICATION ACCURACY ($\overline{OA}$), ITS STANDARD DEVIATION ($s$) AND KAPPA ACCURACY OBTAINED ON TEN RUNS FOR DIFFERENT TRAINING DATA SIZES (INDIAN PINES DATA SET)

| Methods | $\|L\| = 768$ | | | $\|L\| = 868$ | | | $\|L\| = 948$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\overline{OA}$ | $s$ | kappa | $\overline{OA}$ | $s$ | kappa | $\overline{OA}$ | $s$ | kappa |
| SP-EQB | 96.89 | 0.51 | .965 | 97.75 | 0.26 | .974 | 98.12 | 0.32 | .978 |
| SP-MS-ABD | 98.32 | 0.16 | .981 | 98.70 | 0.12 | .985 | 98.94 | 0.14 | .988 |
| SP-CAHT | 97.62 | 0.35 | .973 | 98.07 | 0.31 | .978 | 98.41 | 0.21 | .982 |
| SP-MCLU-ECBD | 98.99 | 0.17 | .989 | 99.21 | 0.20 | .991 | 99.32 | 0.19 | .992 |
| MPM-LBP-BT | **99.64** | 0.17 | **.995** | **99.75** | 0.09 | **.997** | **99.82** | 0.03 | **.998** |
| Proposed | 99.13 | **0.11** | .989 | 99.34 | **0.05** | .992 | 99.44 | **0.02** | .993 |

948 samples in the training set $L$). From this table one can see that the class-wise average classification accuracies obtained by the proposed method are very close to the best results obtained by the literature methods. This again confirms the effectiveness of the proposed AL technique.

TABLE IX
CLASS ACCURACIES (%), AVERAGE OVERALL CLASSIFICATION ACCURACY ($\overline{OA}$) AND ITS STANDARD DEVIATION ($std$), AND AVERAGE KAPPA ($kappa$) ACCURACY OBTAINED ON TEN RUNS (INDIAN PINES DATA SET).

| Methods | SP EQB | SP-MS ABD | SP CAHT | SP-MCLU ECBD | MPM LBP-BT | Proposed |
|---|---|---|---|---|---|---|
| $\|L\|$ | 948 | | | | | |
| **Alfalfa** | **100** | 98.48 | 98.91 | 99.35 | **100** | **100** |
| **Corn-notill** | 92.87 | 97.58 | 95.87 | 98.84 | **100** | 98.87 |
| **Corn-min** | 99.04 | 99.34 | 99.02 | 99.72 | **100** | 99.61 |
| **Corn** | 96.67 | 99.32 | 97.76 | 99.96 | **100** | 99.96 |
| **Grass/Pasture** | 99.05 | 99.77 | 99.48 | 99.96 | **100** | **100** |
| **Grass/Trees** | 99.93 | 99.93 | 99.96 | **100** | **100** | 99.99 |
| **Grass/Pasture-mowed** | 97.86 | 97.14 | 97.14 | 97.14 | **100** | 97.14 |
| **Way-windrowed** | 99.56 | 100 | 99.98 | **100** | **100** | **100** |
| **Oats** | **100** | **100** | **100** | **100** | **100** | **100** |
| **Soybeans-notill** | 94.89 | 96.62 | 96.04 | 97.41 | **98.58** | 97.90 |
| **Soybeans-min** | 99.58 | 99.10 | 98.74 | 99.24 | **99.96** | 99.27 |
| **Soybean-clean** | 99.17 | 99.07 | 97.98 | 99.38 | **100** | 99.56 |
| **Wheat** | 99.85 | 99.95 | 99.80 | **100** | **100** | **100** |
| **Woods** | 99.87 | 99.91 | 99.89 | 99.91 | 99.92 | **99.93** |
| **Bldg-Grass-Tree-Drives** | 99.90 | 99.84 | 99.66 | 99.97 | **100** | 99.97 |
| **Stone-steel towers** | **99.68** | 98.28 | 98.39 | **99.68** | 97.40 | 99.14 |
| $\overline{OA}$ | 98.12 | 98.94 | 98.41 | 99.32 | **99.82** | 99.44 |

### *E. Results: statistical significance test*

In the fourth experiment, for a further comparison between different algorithms, a statistical significance test called $z$-test is utilized [53]. It describes the significance of the difference between two classification results obtained by two different algorithms, which can be calculated as follows:

$$z = \frac{\mu_1 - \mu_2}{\mid \sigma_1^2 - \sigma_2^2 \mid} \tag{6}$$

Where, $\mu_1$ and $\mu_2$ are the mean values of the kappa coefficient obtained by algorithms 1 and 2, respectively and $\sigma_1^2$ and $\sigma_2^2$ are the corresponding variances. If $|z| > 1.96$, the results of two algorithm are assumed to be statistically significant at the 5% significance level.

TABLE X
OBTAINED $Z$-SCORES BETWEEN THE PROPOSED AND THE STATE-OF-THE-ART METHODS FOR ALL THE CONSIDERED DATA SETS.

| Data Sets | SP EQB | SP-MS ABD | SP CAHT | SP-MCLU ECBD | MPM LBP-BT |
|---|---|---|---|---|---|
| **KSC** | 457.65 | 368.18 | 2685.20 | 562.50 | 269.86 |
| **Pavia University** | 60.15 | 2361.10 | 667.05 | 283.33 | 506.36 |
| **Indian Pines** | 1550.80 | 4857.10 | 5162.80 | 875.00 | -1449.30 |

Table X reports the $z$-scores obtained between the proposed technique and the other state-of-the-art methods used for comparison. From the table one can see that except for Indian Pines data set with the MPM-LBP-BT technique, in all the remaining 14 cases the $z$-score obtained between the proposed technique and the state-of-the-art techniques is greater than 1.96. This indicates that the results provided by the proposed technique are statistically significant.

### *F. Results: computation time*

The fifth experiment shows the effectiveness of the different techniques in terms of computational load. All the experiments were carried out on a personal computer (INTEL(R) Core(TM) i5 6500 CPU @3.20 GHz with 4 GB RAM) with the experimental setting (i.e., number of initial training samples, batch size, iteration number, etc.) described in the experiments 1, 2, and 3. Table XI shows the computational time (in minutes) taken by the different techniques for the three considered data sets. From these results one can see that the proposed technique requires significantly less amount of time than the existing spectral-spatial MPM-LBP-BT AL technique. For all the three considered data sets, the MPM-LBP-BT technique takes several hours to complete the AL process. Whereas, the proposed technique needs only few minutes to complete the process. Thus, the MPM-LBP-BT AL technique may not be a reasonable choice for many AL applications. The time taken by the EQB technique is similar to that of the proposed technique. The results reported in Table XI also show that the SP-MS-ABD, the SP-CAHT, and the SP-MCLU-ECBD techniques are faster than the proposed technique. This is because the proposed technique takes some additional time to exploit GAs for selecting informative samples at each iteration of the AL process.

TABLE XI
COMPUTATIONAL TIME (IN MINUTES) TAKEN BY THE DIFFERENT AL METHODS ON THE CONSIDERED DATA SETS.

| Data Sets | SP EQB | SP-MS ABD | SP CAHT | SP-MCLU ECBD | MPM LBP-BT | Proposed |
|---|---|---|---|---|---|---|
| **KSC** | 3 | 1.43 | 1.78 | 1.70 | 371.41 | 6.95 |
| **Pavia University** | 18.46 | 5.81 | 4.83 | 7.16 | 245.91 | 17.15 |
| **Indian Pines** | 14.88 | 3.85 | 5.31 | 4.66 | 60.18 | 15.5 |

### *G. Results: sensitivity analysis*

The final experiment was devoted to analyze the effect of the different parameters used in the proposed technique. The first parameter that may effect the performance of proposed method is the $k$ value associated to the $k$-means clustering. We varied the value of k in the range 400, 500, and 600. Fig. 11 shows the average classification accuracies obtained by the proposed technique for the KSC data set. From this figure one can see that the classification accuracies provided by the proposed technique are not significantly varied within the considered $k$ values. Similar results which are not reported for space constraints are also observed for the other two hyperspectral data sets.

The second parameter that may effect the performance of proposed technique is the $K$ value associated with the $K$-nearest neighbors algorithm used to compute the local density of a region in the feature space. We varied the value of $K$ in the range 5, 10, 15, and 20. Fig. 12 shows the average classification accuracies obtained by the proposed technique on the KSC data set. From the figure one can see that the different values provide very similar results. Similar behavior is also observed for the Pavia University and Indian Pines data sets.

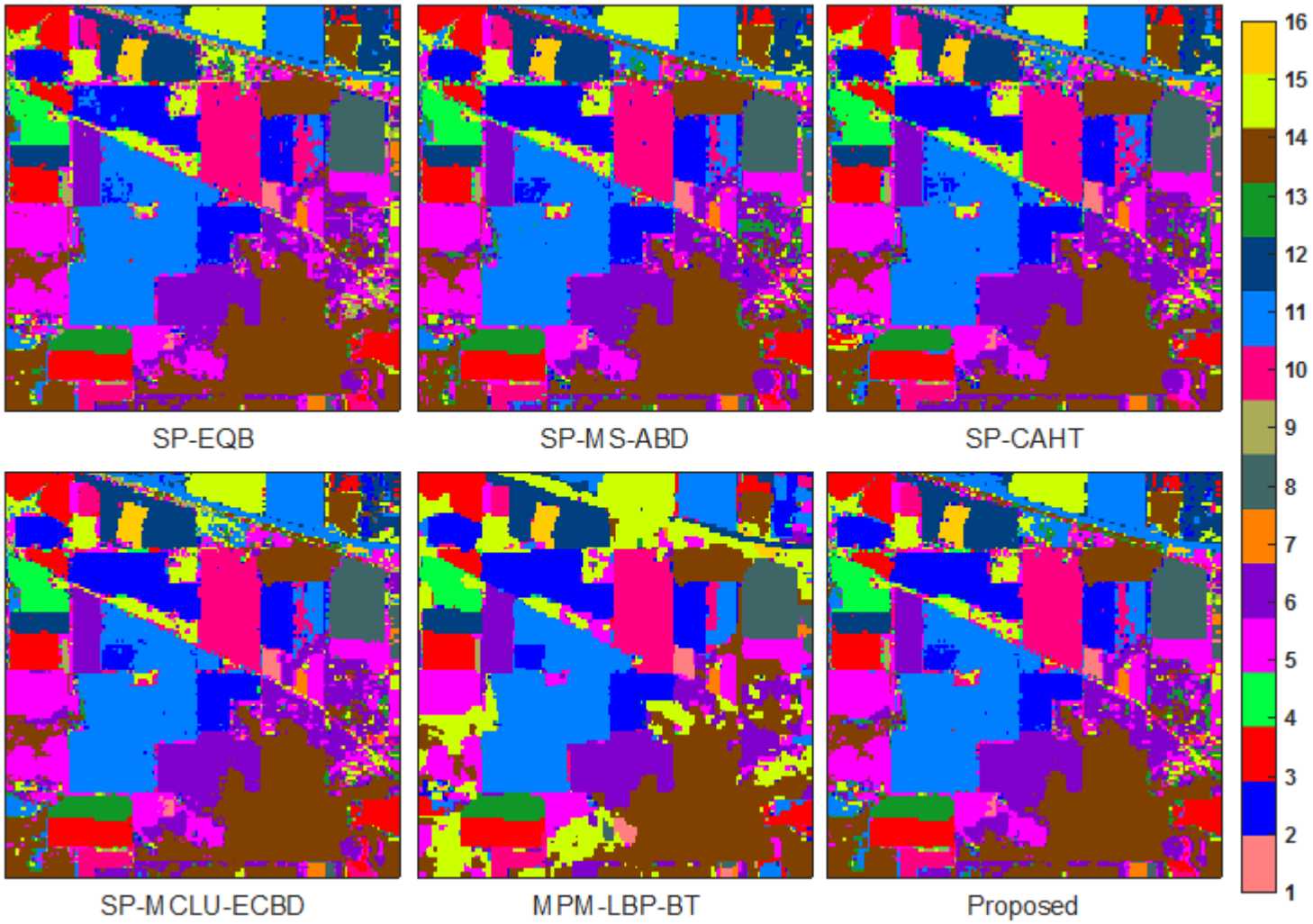


Fig. 10. Classification maps provided by different approaches with 948 labelled samples on the Indian Pines data set.

Finally, we carried out different experiments for assessing the stability of the proposed technique by varying the main parameters of GAs within a wide range. In this regard the population size, the crossover probability and the mutation probability of GAs are varied within the ranges [10 - 40], [0.7 - 0.8] and [0.05 - 0.001], respectively. The results of all these experiments pointed out the low sensitivity of the proposed algorithm to these parameters value within the above defined ranges.

## V. Discussion and Conclusions

In this article a novel technique is presented for classification of HSIs with limited labelled samples. The proposed technique is divided into two phases. Considering the importance of dimensionality reduction and spatial information for the analysis of HSIs, Phase I generates the pattern corresponding to each pixel of HSI by extracting spectral-spatial features. To this end, first, PCA is used to reduce the dimensionality of HSI, then EMPs are exploited. The spectral-spatial features based patterns generated by EMPs are used as input to the Phase II, which performs the classification task with a small number of labelled samples. To this end, a multi-criteria batch mode AL technique is proposed by defining a novel query function that incorporate uncertainty, diversity and cluster assumption criteria. The uncertainty criterion of the proposed query function is defined by exploiting an SVM classifier. The diversity criterion is defined by maximizing the nearest neighbor distances of the selected samples and the cluster assumption criterion is defined by using the properties of $k$-means clustering and $K$-nearest neighbors algorithms. Finally GAs are exploited to select batch of most informative samples by optimizing these three criteria.

To empirically assess the effectiveness of the proposed method, we compared it with five batch mode AL approaches existing in the literature by using three real hyperspectral data sets. By this comparison, we observed that for all the considered data sets, the proposed technique consistently provided better stability with high accuracy. This is due to the integration of the dimensionality reduction, the spectral-spatial feature extraction and the new query function of the AL, which make the proposed technique more robust to the quality of initial labelled samples available. Moreover, the proposed technique is computationally very much less demanding than the one of the existing spectral-spatial information based AL

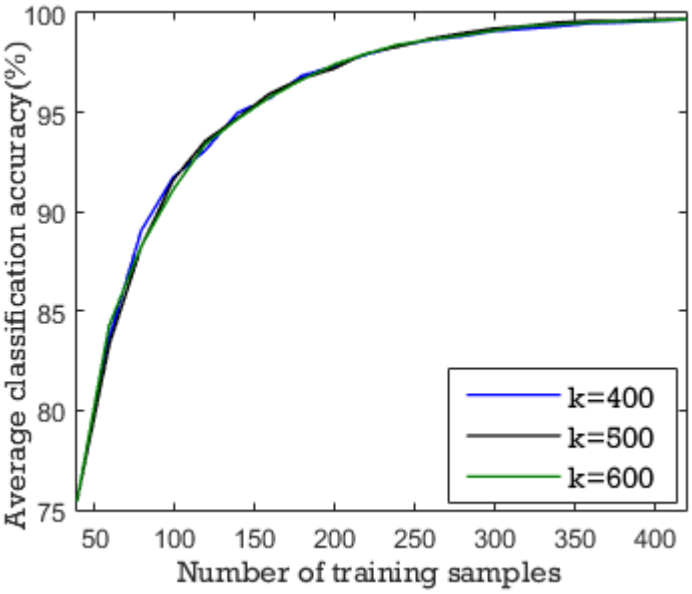


Fig. 11. Average classification accuracy provided by the proposed technique varying the values of $k$ for the $k$-means algorithm (KSC data set).

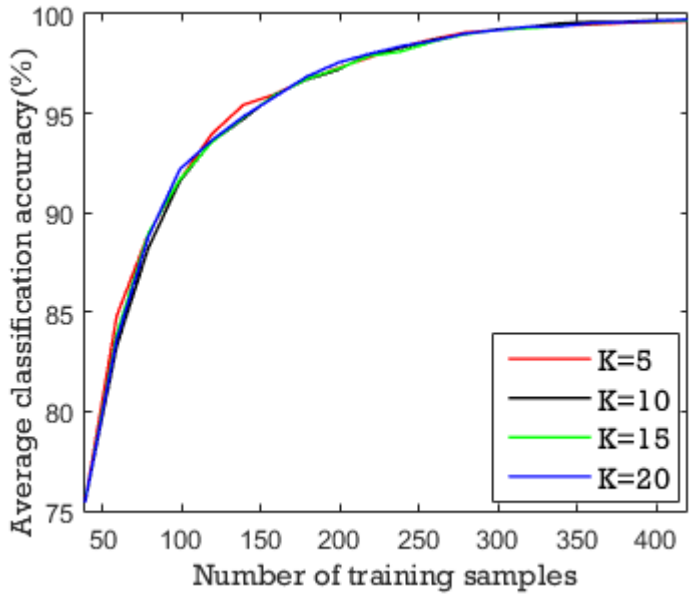


Fig. 12. Average classification accuracy provided by the proposed technique by varying the values of $K$ for the $K$-nearest neighbors algorithm (KSC data set).

technique.

As future developments of this work, we plan to incorporate a multi-objective optimization technique and the use of advanced attribute profile based features in the current AL framework for further improving the classification performance.

## ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees for their constructive criticism and valuable suggestions. Authors would also like to thank the Science and Engineering Research Board, Government of India, under which a project titled Development of Advanced Techniques for the Analysis of Remotely Sensed Images is being carried out at the Department of Computer Science and Engineering, Tezpur University, Assam.

## REFERENCES

[1] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 45–54, 2014.

[2] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Information Theory*, vol. 14, no. 1, pp. 55–63, 1968.

[3] S. B. Serpico and L. Bruzzone, "A new search algorithm for feature selection in hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1360–1367, 2001.

[4] C.-I. Chang and S. Wang, "Constrained band selection for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1575–1585, 2006.

[5] A. Martinez-Uso, F. Pla, J. M. Sotoca, and P. Garcia-Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4158–4171, 2007.

[6] L. Bruzzone and C. Persello, "A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 9, pp. 3180–3191, 2009.

[7] W. Li, S. Prasad, J. Fowler, and L. Bruce, "Locality-preserving dimensionality reduction and classification for hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1185–1198, 2012.

[8] Y. Zhou, J. Peng, and C. Chen, "Dimension reduction using spatial and spectral regularized local discriminant embedding for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 1082–1095, 2015.

[9] H. Huang and M. Yang, "Dimensionality reduction of hyperspectral images with sparse discriminant embedding," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 9, pp. 5160–5169, 2015.

[10] S. Patra, P. Modi, and L. Bruzzone, "Hyperspectral band selection based on rough set," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5495–5503, 2015.

[11] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive SVM for semisupervised classification of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3363–3373, 2006.

[12] G. Camps-Valls, T. B. Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3044–3054, 2007.

[13] M. Marconcini, G. Camps-Valls, and L. Bruzzone, "A composite semisupervised svm for classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 234–238, 2009.

[14] Z. Wang, N. M. Nasrabadi, and T. S. Huang, "Semisupervised hyperspectral classification using task-driven dictionary learning with laplacian regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1161–1173, 2015.

[15] L. Ma, M. Crawford, X. Yang, and Y. Guo, "Local-manifold-learning-based graph construction for semisupervised hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2832–2844, 2015.

[16] D. Tuia, M. Volpi, L. Copa, M. F. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 606–617, 2011.

[17] Z. Wang, B. Du, L. Zhang, L. Zhang, and X. Jia, "A novel semisupervised active-learning algorithm for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3071–3083, 2017.

[18] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *J. Artificial Intelligence Research*, vol. 4, no. 1, pp. 129–145, 1996.

[19] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Machine Learning Research*, vol. 2, no. 1, pp. 45–66, 2002.

[20] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Batch mode active learning with applications to text categorization and image retrieval," *IEEE Trans. Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1233–1248, 2009.

[21] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 1936–1949, 2014.

[22] S. Patra and L. Bruzzone, "A cluster-assumption based batch mode active learning technique," *Pattern Recognition Letters*, vol. 33, no. 9, pp. 1042–1048, 2012.

[23] P. Mitra, B. U. Shankar, and S. K. Pal, "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern Recognition Letters*, vol. 25, no. 9, pp. 1067–1074, 2004.

[24] S. Rajan, J. Ghosh, and M. M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 4, pp. 1231–1242, 2008.

[25] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2218–2232, 2009.

[26] B. Demir, C. Persello, and L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 3, pp. 1014–1031, 2011.

[27] S. Patra and L. Bruzzone, "A batch-mode active learning technique based on multiple uncertainty for SVM classifier," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 3, pp. 497–501, 2012.

[28] ——, "A fast cluster-assumption based active learning technique for classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, no. 5, pp. 1617–1626, 2011.

[29] ——, "A novel SOM-SVM-based active learning technique for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 6899–6910, 2014.

[30] W. Di and M. Crawford, "Active learning via multi-view and local proximity co-regularization for hyperspectral image classification," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 618–628, 2011.

[31] B. Demir, L. Minello, and L. Bruzzone, "Definition of effective training sets for supervised classification of remote sensing images by a novel cost-sensitive active learning method," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1272–1284, 2014.

[32] ——, "An effective strategy to reduce the labeling cost in the definition of training sets by active learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 79–83, 2014.

[33] S. Sun, P. Zhong, H. Xiao, and R. Wang, "Active learning with gaussian process classifier for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1746–1760, 2015.

[34] E. Pasolli, F. Melgani, D. Tuia, F. Pacifici, and W. J. Emery, "SVM active learning approach for image classification using spatial information," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 4, pp. 2217–2233, 2014.

[35] J. Li, J. Bioucas-Dias, and A. Plaza, "Spectral-spatial classification of hyperspectral data using loopy belief propagation and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 844–856, 2013.

[36] S. Sun, P. Zhong, H. Xiao, and R. Wang, "An MRF model-based active learning framework for the spectral-spatial classification of hyperspectral imagery," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 6, pp. 1074–1088, 2015.

[37] Z. Zhang, E. Pasolli, M. M. Crawford, and J. C. Tilton, "An active learning framework for hyperspectral image classification using hierarchical segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 2, pp. 640–654, 2016.

[38] X. Zhou, S. Prasad, and M. M. Crawford, "Wavelet-domain multiview active learning for spatial-spectral hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 9, pp. 4047–4059, 2016.

[39] P. Rigollet, "Generalization error bounds in semi-supervised classification under the cluster assumption," *J. Machine Learning Research*, vol. 8, pp. 1369–1392, 2007.

[40] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901.

[41] J. Serra, *Image analysis and mathematical morphology*. London: Academic Press, 1982.

[42] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, 2005.

[43] M. Dalla Mura, A. Villa, J. A. Benediktsson, J. Chanussot, and L. Bruzzone, "Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 542–546, 2011.

[44] N. Falco, J. A. Benediktsson, and L. Bruzzone, "Spectral and spatial classification of hyperspectral images based on ICA and reduced morphological attribute profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6223–6240, 2015.

[45] X. Huang, X. Han, L. Zhang, J. Gong, W. Liao, and J. A. Benediktsson, "Generalized differential morphological profiles for remote sensing image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 4, pp. 1736–1751, 2016.

[46] W. Liao, J. Chanussot, M. Dalla Mura, X. Huang, R. Bellens, S. Gautama, and W. Philips, "Promoting partial reconstruction for the morphological analysis of very high resolution urban remote sensing images," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 2, pp. 8–28, 2017.

[47] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, 2004.

[48] K. Brinker, "Incorporating diversity in active learning with support vector machines," in *Proc. 20th ICML*, 2003, pp. 59–66.

[49] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley, Singapore, 2001.

[50] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. New York: Addison-Wesley, 1989.

[51] S. Holzwarth, A. Muller, M. Habermeyer, R. Richter, A. Hausold, S. Thiemann, and P. Strobl, "HySens - DAIS 7915/ROSIS imaging spectrometers at DLR," in *Proc. 3rd EARSeL Workshop on Imaging Spectroscopy*, 2003, pp. 3–14.

[52] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machine*, 2001, software available at http://csie.ntu.tw/c̆jlin/libsvm.

[53] G. M. Foody, "Thematic map comparison: evaluating the statistical significance of differences in classification accuracy," *Photogramm. Eng. and Remote Sens.*, vol. 70, pp. 627–634, 2004.