Chapter 2

A spectral-spatial multi-criteria active learning technique for hyperspectral image classification

2.1 Introduction

Hyperspectral images 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 [34]. 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) [115]. 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 [27, 41, 111, 142, 160, 180, 204, 251]. Dimensionality reduction decreases the dimension of hyperspectral data with the help of feature selection (extraction) techniques that select (extract) only 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 can 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 [28, 33, 37, 148, 154, 235]. 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 for classification of HSI with limited labelled samples [229, 236].

The main component of an AL procedure is to design a query function to select the most informative patterns from an unlabelled pool U for labeling. 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 [2, 48, 107, 113, 177, 226, 248]. In the recent years, several AL techniques have been proposed for classification of multispectral and hyperspectral remote sensing images [66–69, 72, 140, 164, 174–176, 178, 190, 216, 217, 227]. Mitra *et al.* [164] 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 the one that is closest to the discriminating hyperplane. In [190], 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 aposteriori probability distribution estimated from the current training set and the training set obtained by including that sample. In [227], 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 [67], Demir *et al.* investigated several SVM-based batch mode AL techniques for the classification of remote sensing images. In [176], a batch mode AL technique based on multiple uncertainty for SVM classifiers is presented. Few cluster assumption based AL techniques are presented in [72, 175, 178]. A cost-sensitive AL method for the classification of remote sensing images is presented in [68] and extended in [69]. This method also includes a cost associated with the accessibility of the unlabelled samples in the query function. An AL technique based on a Gaussian process classifier for hyperspectral image analysis is presented in [216]. 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 [140, 174, 217, 246, 250].

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 chapter we propose a novel AL 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 constructed. The patterns (samples) in 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 the 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 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 are chosen from the m selected samples 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 four other batch mode AL techniques and a spectral-spatial AL technique existing in the literature using four hyperspectral remote sensing data sets.

The rest of this chapter is organized as follows. The proposed active learning technique is presented in Section 2.2. Section 2.3 presents the experimental results obtained on the considered data sets. Finally, Section 2.4 draws the conclusion of this chapter.

2.2 Proposed technique

In this chapter 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. 2-1 shows the block diagram of the proposed framework. The detailed steps of the proposed technique are given in next subsequent subsections.

2.2.1 Phase I: spectral-spatial feature extraction

The classification of an HSI, when a limited number of labelled samples are available, is a challenging task due to the curse of dimensionality problem. Moreover, due to the existence of large number of redundant bands, the distributions of different classes in the original feature space are complex and do not follow the

[1]The cluster assumption says that the boundaries separating the classes should be present in a low density area of the feature space.[193].

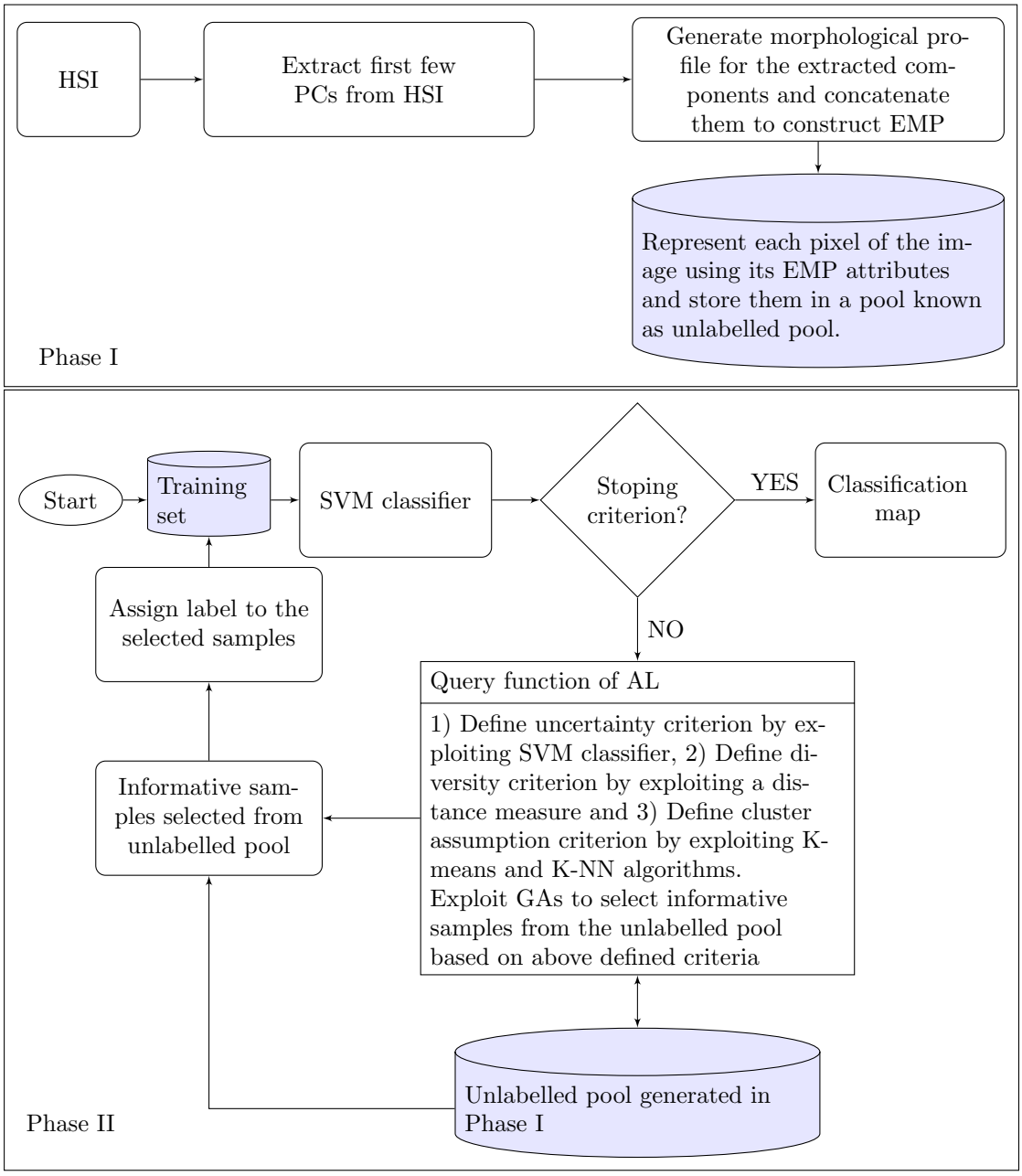

Figure 2-1: Block diagram of the proposed framework.

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 are considered, the class distributions might be much simpler and result in more significant interclass differences. Thus, finding the informative unlabelled samples in the reduced feature space is much easier than the original feature space. Moreover, 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.

Principal component analysis

PCA is an orthogonal transformation technique widely used in feature extraction and data compression [182]. 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 are reduced by keeping only the first l PCs that retain more than 99% of information and discard the rest. More details on PCA is given in Section 1.2.1

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 along with spatial information can reduced the uncertainty level of the patterns 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.

Extended morphological profiles

Mathematical morphology has been successfully applied to images [14, 114, 143, 205]. In the first phase of the proposed method, the dimensionality of HSI is reduced by using PCA and selecting first l PCs. Then, an extended morphological profiles (EMP) is constructed for a hyperspectral image H by using Eq. 1.22 to integrate its spectral and spatial information. The details on morphological operators and construction of an EMP is discussed in Section 1.2.3.

In this phase, morphological opening and closing by reconstruction filters are applied using t different SE on the l PCs extracted out of d dimensional H and $2 \times t$ filtered images are generated for each PC. These filtered images generated for each PC are concatenated along with the original PC to form an EMP with $l(2t+1)$ images containing rich spectral-spatial information to represent the pixels of HSIs. Thus, the patterns corresponding to the pixels of HSI are modeled with $l(2t+1)$ spectral-spatial features.

2.2.2 Phase II: proposed active learning technique for classification of HSIs

After integrating spectral-spatial information for 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 for labeling, the query function of the proposed 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 detail of the proposed technique is given below.

Uncertainty criterion

In this work a one-against-all (OAA) SVM architecture, which involves c binary SVMs (one for each information class), is adopted to define uncertainty criterion as well as to perform the classification task [162]. 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, c binary SVM classifiers are trained with the available labelled samples. After training, c functional distances $f_i(x), i = 1, 2, ..., c$ are obtained, that correspond to the c decision hyperplanes. Then, the classification confidence $CC(x)$ 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. Marginal sampling (MS) is the first strategy, in which the CC of each unlabelled sample is computed by considering the smallest distance among the c decision hyperplanes [164]. The second strategy, which is also used in this work, is based on the multi-class label uncertainty (MCLU) [67]. MCLU computes the CC of each unlabelled sample $x \in U$ by considering the difference between the first and second largest distance values to the hyperplanes using Eq. 1.12.

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, it means that the model has high confidence in assigning the sample x into the class corresponding to the maximum distance. On the contrary, if $CC(x)$ is small, the classifier model has low confidence in assigning the sample x to any of the class.

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 [23, 67, 227]. In this work a simple criterion that maximizes the distance between the 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:

$$\max \left\{ \sum_{i=1}^{h} \min_{\substack{i \neq j \\ j=1,...,m}} \{d(x_i, x_j)\} \right\} \tag{2.1}$$

where $d(x_i, x_j)$ is the euclidean distance between the samples x_i and x_j. The h samples selected using Eq. (2.1) are diverse from each other since the criterion maximize the distance between each sample with its nearest sample.

Cluster assumption criterion

The cluster assumption property states that the boundaries separating the classes should be present in a low density area of the feature space. Thus, the patterns belonging to the 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 an 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 [76]. 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 $Cl_1, Cl_2, ..., Cl_k$ 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 Cl_i belongs can be computed as follows:

$$den(Cl_i) = \frac{1}{K} \sum_{x_i \in K-NN(\mu_i)} d(\mu_i, x_i) \tag{2.2}$$

where $K - NN(\mu_i)$ represents the K neighbor patterns that are nearest to the cluster representative μ_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:

$$den(x_j) = den(Cl_i), \text{where } x_j \in Cl_i \tag{2.3}$$

According to the cluster assumption, the patterns having higher $den(x_j)$ values have higher probability to be in a low-density region in the feature space as compared to the patterns having lower $den(x_j)$ values. Thus, the density computed

by Eq. 2.3 can be used to evaluate the cluster assumption property in the AL query.

Selecting informative samples using GAs

In this section a query strategy for AL based on the above-defined criteria is presented by exploiting GAs [99]. At each iteration of AL, first the m samples from U that have the lowest classification confidence computed using Eq. 1.12 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.

Representation of chromosome: Each chromosome represents the h samples by a sequence of binary numbers. If a sample is represented by s bits, the length of a chromosome representing 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.

Initialization of the Population: 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:

$$F(x_1, x_2, ..., x_h) = \\ \frac{1}{h}\sum_{i=1}^{h} CC(x_i) - \frac{1}{h}\left\{\sum_{i=1}^{h} \min_{\substack{i \neq j \\ j=1,...,m}} \{d(x_i, x_j)\}\right\} - \frac{1}{h}\sum_{i=1}^{h} den(x_i) + Pen \tag{2.4}$$

Here h ($h < m$) informative samples are chosen from the m uncertain samples (obtained by using the uncertainty criterion defined in Eq. 1.12) 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 Eq. 1.12), the average minimum neighbor distance (using the diversity criterion defined in Eq. 2.1) and the average density (using the cluster

assumption criterion defined in Eq. 2.3) of the h samples represented by a chromosome, respectively. If a sample appear multiple times in a chromosome, the parameter Pen 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 (2.4) 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. Here the 'stochastic uniform' selection strategy is adopted.

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

Mutation: With a fixed probability, each chromosome undergoes mutation. Given a chromosome in the population, the value at a bit position (or gene) is flipped in mutation.

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 3 provides the details of the proposed AL technique.

2.3 Experimental Results

2.3.1 Design of experiments

In order to show the potential of the proposed technique all the four hyperspectral data sets described in Appendix A 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) [227]; ii) the marginal sampling with angle based diversity

Algorithm 2 Proposed active learning technique

Phase I

1: Apply PCA to HSI and select first l PCs.
2: Obtain MP of l component images generated for 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 with its EMP features.

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 Eq. 2.2 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 Eq. 2.3.
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 Eq. 1.12.
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 (2.4).
7: Assign labels to the h selected samples and include them into the training set.
8: **until** the stoping criterion is satisfied.

(MS-ABD) [23]; iii) the cluster assumption with histogram thresholding (CAHT) [176]; and iv) the multiclass label uncertainty with enhanced cluster based diversity (MCLU-ECBD) [67]. 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 employs bagging for forming a committee of classifier and considers the maximum disagreement between them to directly select the h most uncertain samples. 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 is also compared with an existing spectral-spatial information based state-of-the-art AL technique referred as MPM-LBP-BT technique [140]. 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 2.2 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 AL techniques presented in this chapter have been implemented in Matlab (R2015a). OAA SVM with radial basis function (RBF) kernels has been implemented by using the LIBSVM library [38]. The SVM parameters $\{C, \sigma\}$ (the regularization parameter and the spread of the RBF kernel) for all the data sets are derived by applying a grid search with a five-fold cross-validation technique. During the active learning iterations these values were kept same for simplicity.

2.3.2 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 (see Appendix A.1). For this experiment, a total of $T = 5211$ labelled samples (see Table A.1) 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. 20 samples were selected from U at each iteration of AL for labeling and the process was iterated 19 times resulting in 419 samples in the training set L. The active learning process was repeated for 10 trials with different initial labelled samples for reducing the random effect on the results.

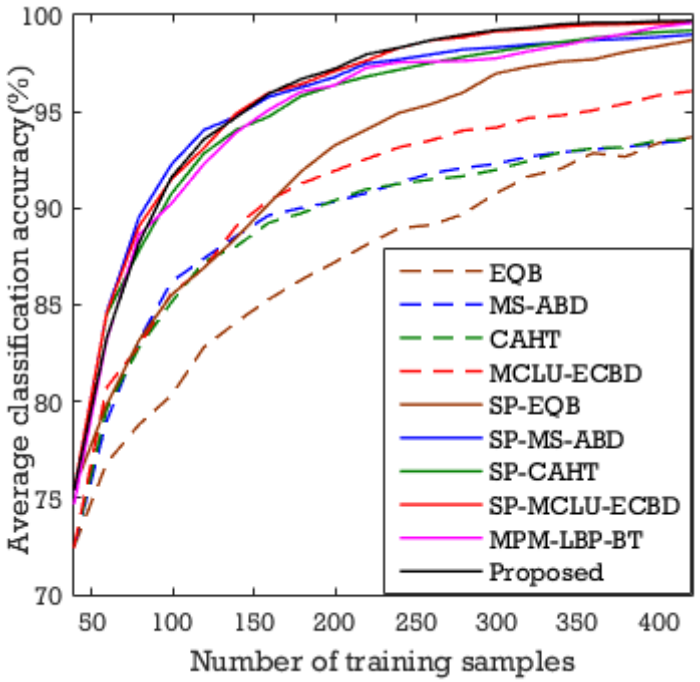

Figure 2-2: Average classification accuracy over ten runs versus the number of training samples provided by the different methods (KSC data set).

Fig. 2-2 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 by 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 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 2.1: 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	$\|L\| = 239$			$\|L\| = 339$			$\|L\| = 419$		
	$\overline{OA}$	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 2.1 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 2.2 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 close to the best

Table 2.2: 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**

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. 2-3 shows the classification maps obtained by the different AL techniques where the map obtained for the proposed method is more regularized.

2.3.3 Results: University of Pavia data set

In order to assess the effectiveness of the proposed technique, the second experiment is carried out considering the University of Pavia data set (see Appendix A.2). For this experiment, $T = 42776$ labelled samples (see Table A.2) 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. Also in this one the active learning process was repeated for 10 trials with different initial labelled samples for reducing the random effect on the results

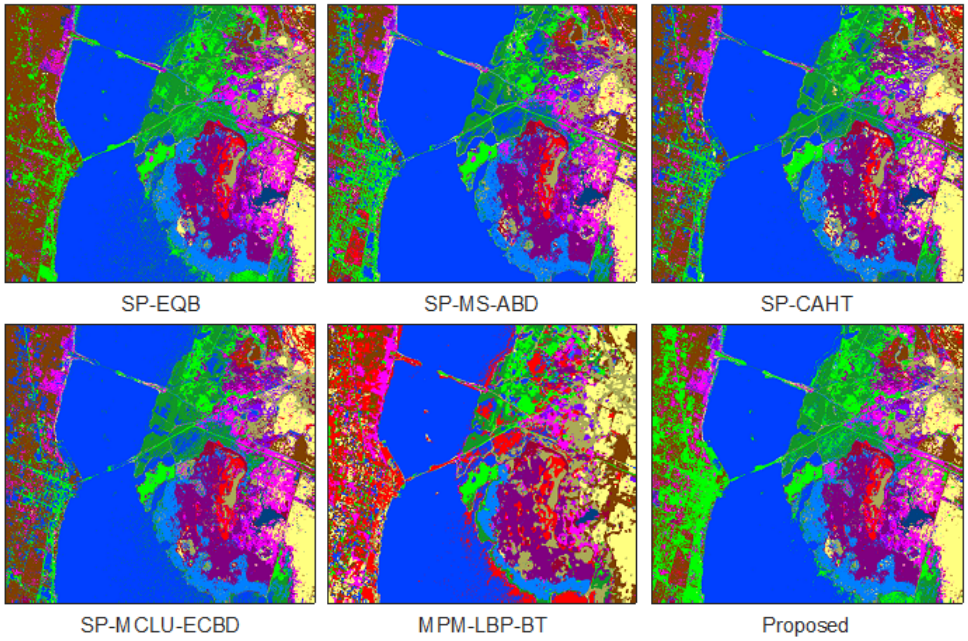

Figure 2-3: Classification maps provided by different approaches with 419 labelled samples on the KSC data set.

Fig. 2-4 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 University of Pavia data set, among the six spectral-spatial AL techniques, the MPM-LBP-BT technique provides worst classification results.

Table 2.3: Average overall classification accuracy ($\overline{OA}$), its standard deviation (s) and kappa accuracy obtained on ten runs for different training data sizes (University of Pavia 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**

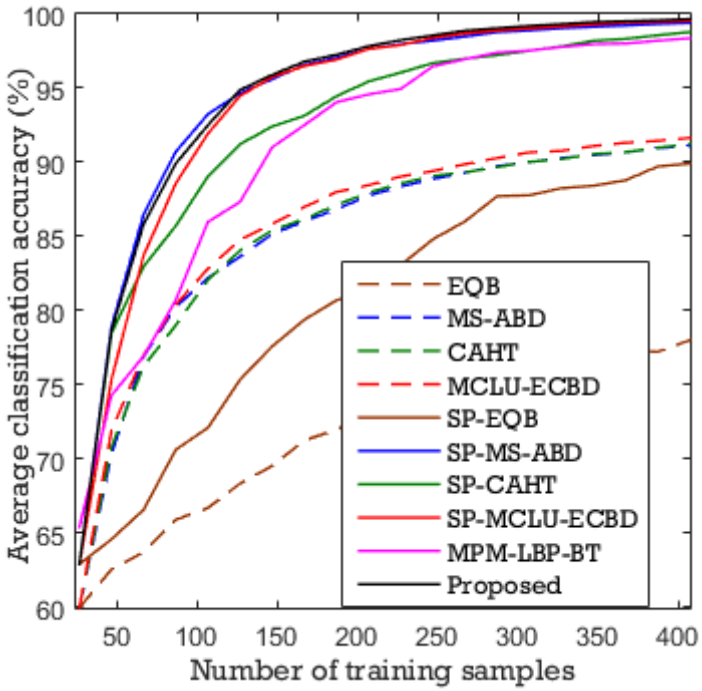

Figure 2-4: Average classification accuracy over ten runs versus the number of training samples provided by the different methods (University of Pavia data set).

Table 2.3 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 2.4 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. 2-5 shows the classification maps obtained by the different AL techniques for visual analysis.

Table 2.4: Class wise average classification accuracies (%) obtained on ten runs (University of Pavia 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**

2.3.4 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 (see Appendix A.3). A total of $T = 10249$ labelled samples (see Table A.3) 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 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. 2-6 shows the average overall classification accuracies provided by

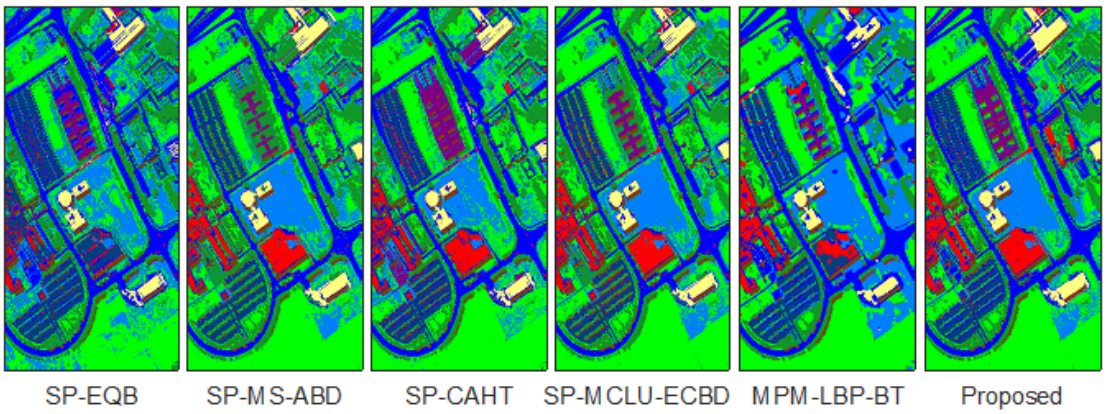

Figure 2-5: Classification maps provided by different approaches with 407 labelled samples on the University of Pavia data set.

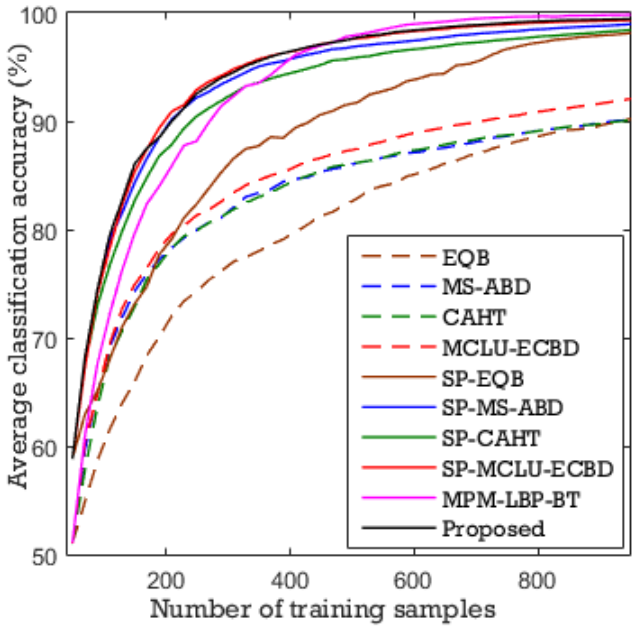

Figure 2-6: Average classification accuracy over ten runs versus the number of training samples provided by the different methods (Indian Pines data set).

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 2.5: 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

Table 2.5 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 University of Pavia data sets. Table 2.6 shows the average class-wise accuracies (%) obtained by different AL techniques after completing 45 iterations (i.e., 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. For visual analysis Fig. 2-7 shows the classification maps obtained by the different AL techniques.

Table 2.6: Class wise average classification accuracies (%) 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

2.3.5 Results: University of Houston data set

In order to assess the effectiveness of the proposed technique, the fourth experiment is carried out on the University of Houston data set (see Appendix A.4). A total of $T = 15029$ labelled samples (see Table A.4) are considered as a test set TS. For this experiment, first only 45 samples (three samples from each class) are randomly selected from T as training set L, and the remaining 14984 are stored in the unlabelled pool U. At each iteration of AL, 20 samples are selected from

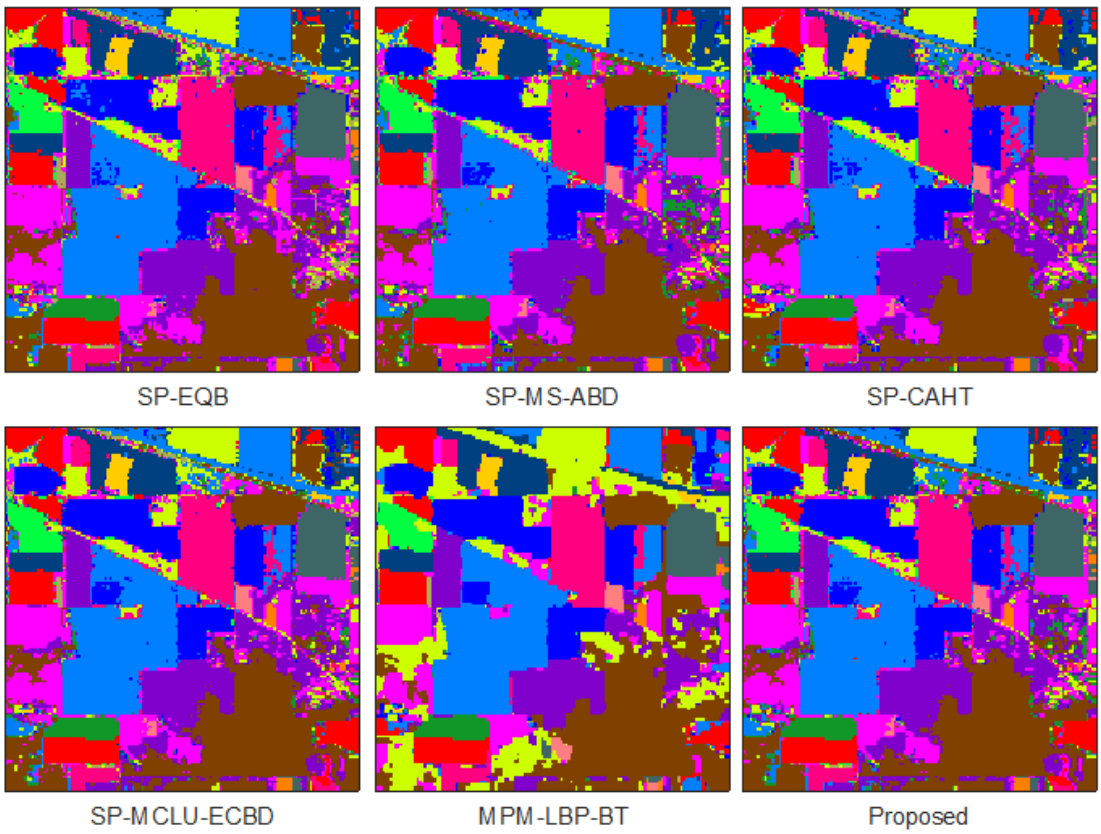

Figure 2-7: Classification maps provided by different approaches with 948 labelled samples on the Indian Pines data set.

U for labeling and the process is iterated 45 times resulting in 945 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. 2-8 shows the average overall classification accuracies provided by the different methods versus the number of samples included into the training set for University of Houston 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 4%) 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 the proposed technique provided better results than the existing MPM-LBP-BT technique.

Table 2.7 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 highest classification accuracy among the considered AL techniques. Table 2.8 shows the average class-wise accuracies (%)

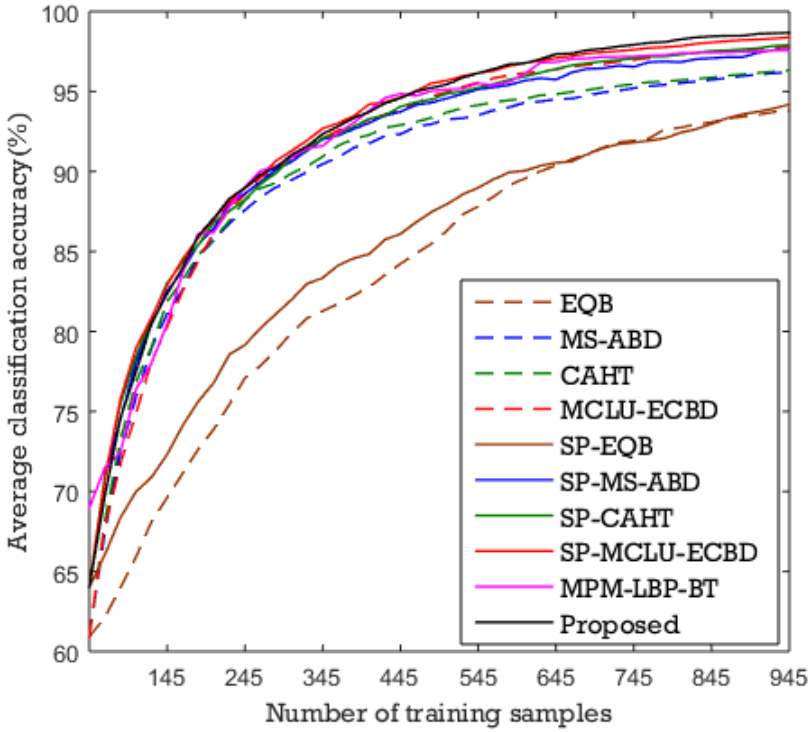

Figure 2-8: Average classification accuracy over ten runs versus the number of training samples provided by the different methods (University of Houston data set).

Table 2.7: Average overall classification accuracy ($\overline{OA}$), its standard deviation (s) and kappa accuracy obtained on ten runs for different training data sizes (University of Houston Data Set)

Methods	$\|L\| = 765$			$\|L\| = 865$			$\|L\| = 945$		
	$\overline{OA}$	s	kappa	$\overline{OA}$	s	kappa	$\overline{OA}$	s	kappa
SP-EQB	91.93	1.38	0.913	93.33	1.16	0.928	94.19	1.00	0.937
SP-MS-ABD	96.84	0.61	0.966	97.15	0.67	0.969	97.63	0.64	0.974
SP-CAHT	97.19	**0.26**	0.970	97.57	**0.24**	0.974	97.90	**0.12**	0.977
SP-MCLU-ECBD	97.72	0.78	0.975	98.15	0.53	0.980	98.38	0.41	0.983
MPM-LBP-BT	97.24	0.48	0.970	97.44	0.39	0.972	97.57	0.37	0.974
Proposed	**98.04**	0.44	**0.979**	**98.48**	0.35	**0.984**	**98.67**	0.34	**0.986**

obtained by different AL techniques after completing 45 iterations (i.e., 945 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 consistently good for all the classes. This again confirms the effectiveness of the proposed AL technique. For visual analysis Fig. 2-9 shows the classification maps obtained by the different AL techniques.

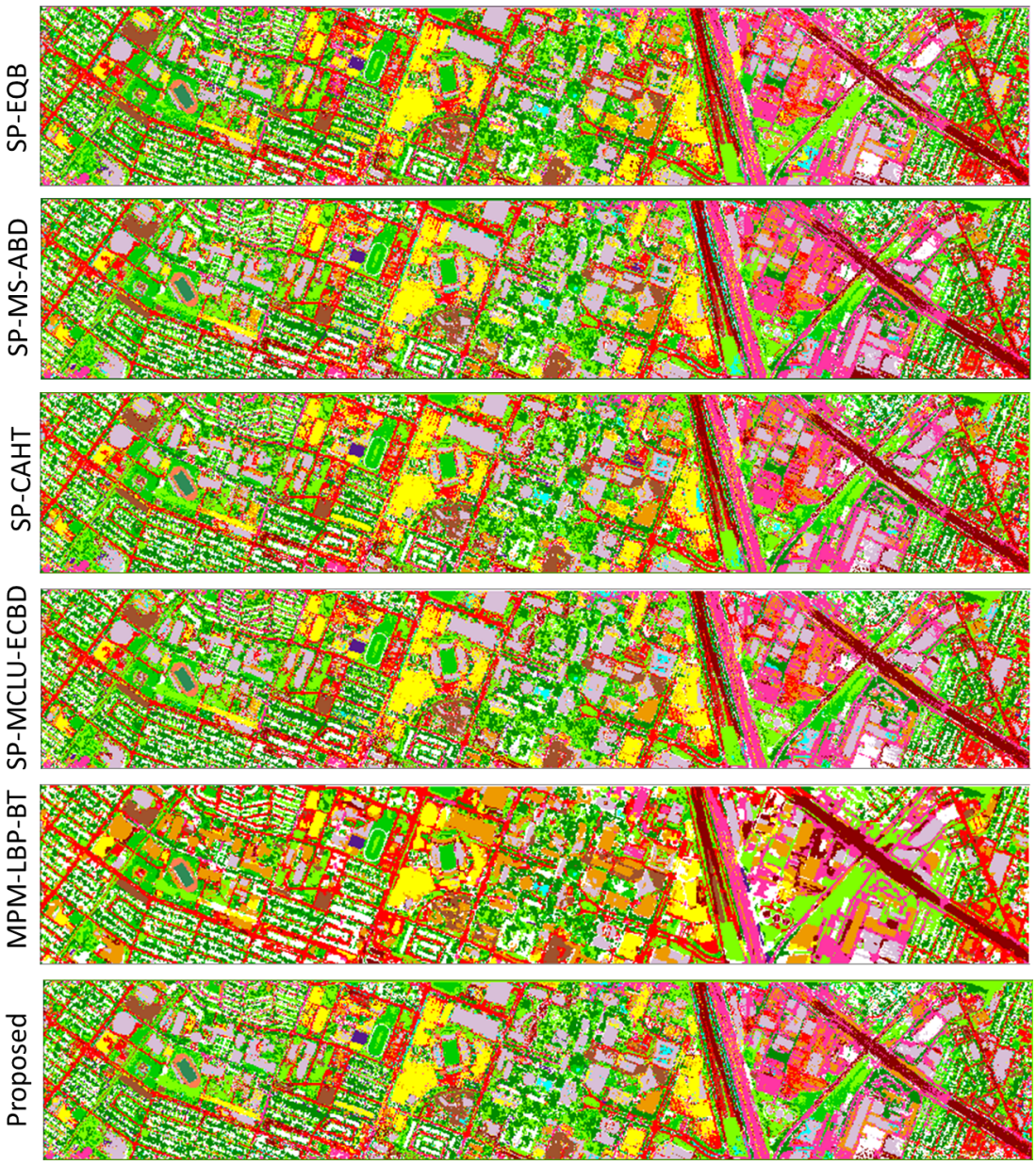

Figure 2-9: Classification maps provided by different approaches with 945 labelled samples on the University of Houston data set.

Table 2.8: Class wise average classification accuracies (%) obtained on ten runs (University of Houston Data Set)

Methods	SP EQB	SP-MS ABD	SP CAHT	SP-MCLU ECBD	MPM LBP-BT	Proposed
	$\|L\|$ =**945**					
Grass-Healthy	87.55	99.20	99.39	99.19	98.03	**99.63**
Grass-Stressed	94.48	96.24	98.06	94.24	**100.00**	96.27
Grass-Synthetic	99.81	99.97	99.86	99.97	**100.00**	99.97
Tree	94.53	99.73	99.20	99.71	**100.00**	99.69
Soil	98.86	99.93	99.76	**100.00**	**100.00**	**100.00**
Water	97.32	99.14	99.57	**99.94**	98.07	99.91
Residential	96.14	99.01	98.80	99.90	**100.00**	99.67
Commercial	87.71	93.97	98.00	97.86	77.72	**98.31**
Road	92.81	95.49	95.15	**97.64**	95.64	97.50
Highway	90.48	97.50	97.52	97.27	**100.00**	98.06
Railway	95.60	96.27	95.22	97.87	**100.00**	97.90
Parking Lot 1	95.76	97.13	96.51	97.60	**100.00**	97.75
Parking Lot 2	92.67	92.58	92.96	**97.89**	95.29	97.78
Tennis Court	98.74	99.93	99.74	**100.00**	**100.00**	**100.00**
Running Track	99.95	99.91	99.88	99.97	**100.00**	99.97
$\overline{OA}$	94.19	97.63	97.9	98.38	97.57	**98.67**

2.3.6 Results: statistical significance test

In the fifth experiment, for a further comparison between different algorithms, a statistical significance test called z-test is utilized [85]. It describes the significance of the difference between two classification results obtained by two different algorithms, which can be calculated as follows:

$$z = \frac{\mu_1 - \mu_2}{\mid \sigma_1^2 - \sigma_2^2 \mid} \tag{2.5}$$

Where, μ_1 and μ_2 are the mean values of the kappa coefficient obtained by algorithms 1 and 2, respectively and σ_1^2 and σ_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 2.9 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 in all the 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.

Table 2.9: 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
University of Pavia	60.15	2361.10	667.05	283.33	506.36
Indian Pines	1550.80	4857.10	5162.80	875.00	1449.30
University of Houston	5.07	3.57	7.61	5.52	51.64

2.3.7 Results: computation time

The sixth 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, 3 and 4. Table 2.10 shows the computational time (in minutes) taken by the different techniques for the four 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 four 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 2.10 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 2.10: 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
University of Pavia	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
University of Houston	9.5	1.33	2.74	1.89	1841.8	6.18

2.3.8 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. 2-10 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 are also observed for the other data sets.

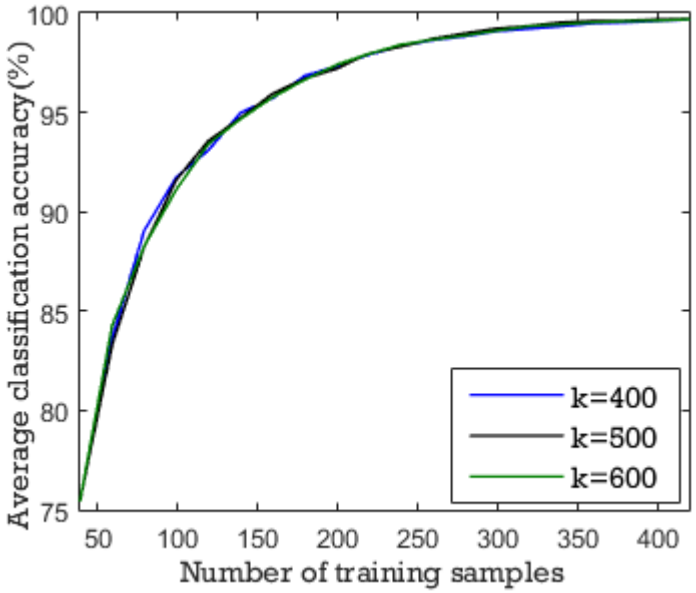

Figure 2-10: Average classification accuracy provided by the proposed technique varying the values of k for the k-means algorithm (KSC data set).

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. 2-11 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.

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

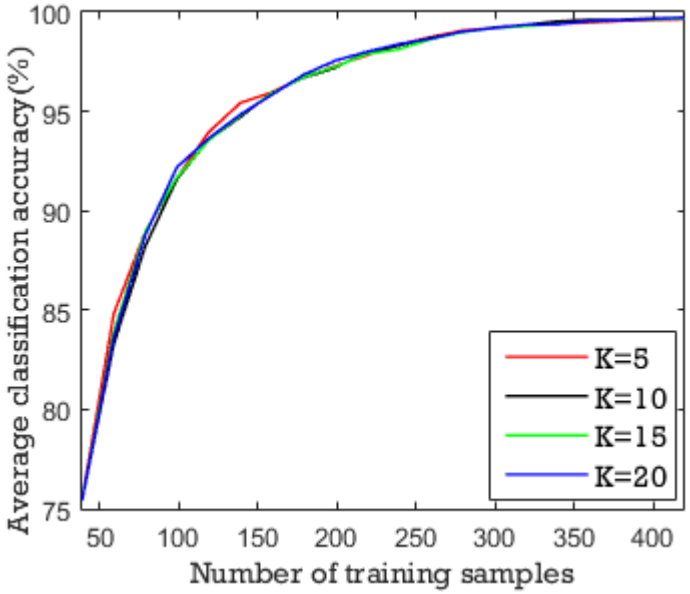

Figure 2-11: Average classification accuracy provided by the proposed technique by varying the values of K for the K-nearest neighbors algorithm (KSC data set).

defined ranges.

2.4 Discussion and conclusions

In this chapter a novel AL 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 incorporates 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 com-

pared it with five batch mode AL approaches existing in the literature by using four 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 technique.