Title: A novel SOM-SVM based active learning technique for remote sensing image classification

This paper appear in: IEEE Transaction on Geoscience and Remote Sensing

Date of Publication: 2014

Author(s): Swarnajyoti Patra and Lorenzo Bruzzone

Volume: 52, Issue: 11

Page(s): 6899-6910

DOI: 10.1109/TGRS.2014.2305516

A novel SOM-SVM based active learning technique for remote sensing image classification

Swarnajyoti Patra and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—In this article a novel iterative active learning technique based on self-organizing map (SOM) neural network and support vector machine (SVM) classifier is presented. The technique exploits the properties of the SVM classifier and of the SOM neural network to identify uncertain and diverse samples, to include in the training set. It selects uncertain samples from low-density regions of the feature space by exploiting the topological properties of the SOM. This results in a fast convergence also when the available initial training samples are poor. The effectiveness of the proposed method is assessed by comparing it with several methods existing in the literature using a toy data set and a color image as well as real multispectral and hyperspectral remote sensing images.**

***Index Terms*—Active learning, support vector machine, self-organizing map, hyperspectral imagery, multispectral imagery, remote sensing.**

I. Introduction

In supervised techniques, the classification accuracy depends on the quality of labeled patterns used for training. The collection of informative labeled samples is usually expensive and time consuming. When considering remote sensing image classification problems, we may have several millions of unlabeled patterns (pixels); thus the manual selection of the training samples (usually carried out according to pre-defined sampling strategies) is a complex process and often introduces redundancy into the training set. In order to both reduce the cost of labeling and optimize the performance of the classifier, the training set should be as small as possible by avoiding redundant samples and including only most informative patterns (which have the highest training performance). The active learning approach addresses this problem. Like in a human-machine interaction scenario, the learning process repeatedly queries unlabeled samples to select the most informative patterns for the considered learning technique. Then it updates the training set on the basis of a supervisor who attributes the labels to the selected unlabeled samples. Thus, the unnecessary labeling of noninformative samples is avoided greatly reducing the labeling cost, while increasing the quality of the training set.

Active learning techniques are widely used in pattern recognition literature [1]–[9]. All the methods differ only in their query function, which is the core of the active learning process. The query function can be designed by taking into account an uncertainty criterion followed by a diversity criterion. The uncertainty criterion is associated to the confidence of the supervised algorithm in correctly classifying the considered sample, while the diversity criterion aims at selecting a set of unlabeled samples that are as diverse as possible in the feature space, thus reducing the redundancy among the samples selected at each iteration. Many active learning methods at each iteration chose either i) single [1]–[3] or ii) multiple [4], [5] informative samples for labeling by considering only the uncertainty criterion. The first approach is inefficient since the classifier needs to be retrained after adding only a single sample into the training set. The second approach can be inefficient due to the high possibility of selecting redundant samples. To mitigate both the above-mentioned problems, some active learning techniques query a batch of unlabeled samples at each iteration by considering both uncertainty and diversity criteria [6], [7].

In this paper, we focus on classification of remote sensing images. However the proposed method is general and can be used in any classification problem. Active learning methods have been increasingly considered in remote sensing image classification only in very recent years [10]–[18]. Some of them select single [10], [11] and some multiple [12]–[14] samples at each iteration of the active learning process by considering only the uncertainty criterion. Mitra et al. [10] presented an active learning technique that selects the most uncertain sample closest to the current separating hyperplane of an SVM classifier. Rajan et al. [11] presented an active learning method that chooses the unlabeled sample that maximizes the information gain between the a posterior probability distribution estimated from the current training set and the one obtained by including that sample into it. The information gain is measured by the Kullback-Leibler divergence. In [12], Tuia et al. presented a technique that selects multiple samples at each iteration of the active learning process with the help of the entropy query-by-bagging algorithm. The samples that have maximum disagreement among the committee of learners are considered as most uncertain. Recently, few techniques were developed that introduce cluster assumption to select the most uncertain samples [13], [14]. In [13], we proposed a simple cluster assumption based method that selects the samples to be labeled from low-density regions of one-dimensional SVM output space. In [14], Di and Crawford investigated a coregularization method that incorporates the inconsistency from both the local proximity and multiview perspectives, whereby the local proximity is enforced and measured on the spatial/spectral generated manifold space. All these methods select a set of most uncertain samples at each iteration of the active learning process without incorporating a diversity criterion, thus possibly introducing redundancy into the selected samples. Other active learning methods mitigate this problem by incorporating a diversity criterion in the sample selection process [12], [15], [16]. All these methods follow

two steps, the uncertainty step and the diversity step. In the uncertainty step, the m $(m > 1)$ most uncertain samples are selected using a given uncertainty criterion. Then in the diversity step, h $(1 < h < m)$ samples are selected among the m most uncertain samples by applying a diversity criterion. In [12], Tuia et. al. presented a method that selects the m most uncertain samples which are closest to the SVM decision hyperplanes; then the h samples among the m which are closest to distinct support vectors are chosen for labeling. In [15], Demir et al. investigated several SVM-based batch mode active learning techniques by incorporating different diversity measures. In [16], we developed a batch-mode active learning technique based on multiple uncertainty for SVM classifiers. Recently, Demir et al. proposed cost-sensitive active learning methods for the classification of remote sensing images[19], [20]. They modeled the query function of the active learning not only considering the uncertainty and diversity criteria, but also including explicitly the labeling cost information. A survey of active learning methods in remote sensing literature is presented in [21].

In remote sensing classification problems, the collection of labeled samples for the labeling of selected samples at each iteration of active learning can be obtained according to the following: 1) ground survey, which is costly and time consuming; 2) photointerpretation (expert interpretation of the image), which is cheaper and faster; or 3) hybrid solutions, where both photointerpretation and ground surveys are used. The choice of the labeling strategy depends on the considered problem and image type. For example, for very high resolution (VHR) images, the labeling of a particular sample can be usually easily obtained by photointerpretation. When medium (or low) resolution multispectral images and hyperspectral data are considered, the land-cover classes are characterized usually on the basis of their spectral signatures. In these cases, the visual analysis of different false color compositions (i.e., photointerpretation) often is not sufficient to predict the appropriate label of a particular sample. Thus, ground survey is necessary for the labeling of samples. According to these example we can conclude that depending on both the type of classification problem and the considered image type, the cost and time associated to the labeling process significantly changes. Several iterations of the labeling step in the active learning strategy can be carried out where photointerpretation is possible. On the contrary, in cases where ground surveys are necessary, only few iterations of the labeling step are possible[15]. These iterations should be carried out minimizing the cost of labeling (see [19], [20] for more details on the cost issue).

In this paper we present a novel batch mode active learning technique based on self-organizing map (SOM) neural networks [22] and support vector machine (SVM) classifiers [23], [24]. The proposed technique exploits the cluster assumption to find and select the most informative samples among those selected by applying both uncertainty and diversity criteria at each iteration of the active learning process. 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 [25]. In our active learning method, first, a SOM network is trained in an unsupervised way with the available unlabeled patterns or with a sub-sampled set of unlabeled patterns for limiting the learning time. After training, we compute the average distance of each neuron in the output layer to its neighbor neurons using their corresponding weight vectors. Under the assumption that SOM preserves the topological property of the input patterns, the neurons mapping the samples that belong to low-density regions of the input space have larger average neighbor distance than the neurons mapping the samples that belong to high-density regions. In other words, according to the cluster assumption we can say that neurons which have higher average neighbor distance have a high probability to map boundary samples. At convergence of the SOM processing we start the iterative active learning procedure. Initially we train the SVM classifier with the available labeled samples. After training, we compute the confidence of correct classification of each unlabeled sample with the help of the trained SVM. Then the h_1 unlabeled samples that both have the lowest classification confidence and are mapped into distinct neurons of the SOM are selected. This allows us to select the h_1 most uncertain samples which are diverse from each other. Then a batch of h $(h < h_1)$ samples from the selected h_1 samples are chosen that correspond to the SOM mapping neurons having the highest average neighbor distance. This allows us to incorporate the cluster assumption property to select the most informative samples for labeling. Thus, the proposed technique can easily use the cluster assumption and a diversity criterion in the sample selection process by exploiting the properties of the SOM neural network. The main advantage of using the cluster assumption is that in this way we can locate with higher precision relevant training samples close to the decision boundary between classes also when biased initial training samples are considered.

The proposed method is compared with several other active learning methods existing in the remote sensing literature by using one toy data set, a color image, as well as two real remote sensing data sets made up of a multispectral image and a hyperspectral image. Experimental results show the effectiveness of the proposed method.

The rest of this paper is organized as follows. The proposed SOM-SVM based active learning approach is presented in Section II. Section III provides the description of the four data sets used for experiments. Section IV presents experimental results obtained on the considered data sets. Finally, Section V draws the conclusion of this work.

II. Proposed method

We present a novel batch mode active learning technique based on SOM neural networks and SVM classifiers for solving multiclass classification problems. Before presenting the proposed technique, we briefly recall the main concepts associated with both SOM neural networks and SVM classifiers. The reader is referred to [22] and [23], [26] for more details on the SOM and the SVM approaches, respectively.

A. Self-organizing map neural network

The self-organizing map is a popular artificial neural network algorithm based on unsupervised learning. The SOM is able to project high-dimensional data into a lower dimension that can be useful for analyzing the patterns in the input space [27], [28]. Fig. 1 presents the architecture of a SOM neural network. The network consists of an input and an output layer. The number of neurons in the input layer is equal to the dimension of the feature input vector. The output layer consists in a regular 2D grid of neurons called map. The neurons of the map can be arranged either on a rectangular or a hexagonal lattice, where each neuron in the map is connected with all the neurons in the input layer by using a weight vector.

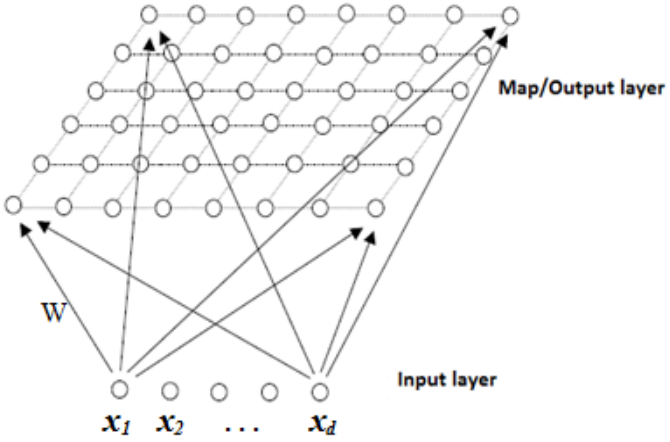


Fig. 1. Architecture of a SOM neural network.

The SOM algorithm is based on the competitive learning concept. When a training sample (which does not include information on the class label) is fed in input to the network, a metric distance is computed for all weight vectors. The neuron of the map with the weight vector most similar to the input pattern is called the best matching unit (BMU). The weights of the BMU and its neighboring neurons are then adjusted towards the input pattern. The magnitude of the change decreases with time and with distance from the BMU. After training, the map produced by the SOM algorithm preserves the topological property of the input patterns, i.e., weight vectors which are neighbors in the input space are mapped onto neighboring neurons of the map.

In greater detail, let $X \in \Re^d$ and $W \in \Re^d$ be the set of input and weight vectors in a d-dimensional space, respectively. Let each neuron k of the map have an associated weight vector $w_k \in W$. The initial values for the weight vectors can be set randomly. The SOM algorithm either follows sequential or batch training to update the weight vectors [22].

Sequential training: In this method the weight vectors are updated immediately upon the presentation of an input pattern. Thus the sequential SOM algorithm can be formalized as follows:

Batch training: In this method the whole training set is analyzed at once and only after this analysis the map is updated considering the effects of all the samples. The new weight vectors are computed as follows:

Algorithm 1 Sequential SOM training

1: Randomly select a sample x_i from the training set X.
2: Find the corresponding BMU, denoted as c_i, as follows:

$$c_i = \arg \min_{k=1,2,...,|W|} \{\| x_i - w_k(t) \|^2\}. \quad (1)$$

3: Update the weight vector of neuron $k(k = 1, 2, ..., |W|)$ as follows:

$$w_k(t+1) = w_k(t) + \eta(t) h_{c_i k}(t)[x_i - w_k(t)] \quad (2)$$

where t denotes time, $0 < \eta(t) < 1$ is the learning rate parameter, and the scalar multiplier $h_{c_i k}(t)$ is the neighborhood kernel around the winner neuron c_i.
4: Repeat from step 1 for all training patterns $x_i(i = 1, 2, ..., |X|)$, completing one epoch.
5: Decrease the value of neighborhood kernel and learning rate.
6: Repeat from step 1 until convergence criterion is met.

$$w_k(t+1) = \frac{\sum_{i=1}^{|X|} h_{c_i k}(t) x_i}{\sum_{i=1}^{|X|} h_{c_i k}(t)} \quad (3)$$

The batch SOM algorithm can be formalized as follows:

Algorithm 2 Batch SOM training

1: Find the BMU for an input vector x_i using (1).
2: Accumulate numerator and denominator of (3) for all neurons.
3: Repeat from step 1 for all training patterns $x_i(i = 1, 2, ..., |X|)$, completing one epoch.
4: Update neuron weights with (3).
5: Decrease the value of neighborhood kernel.
6: Repeat from step 1 until convergence criterion is met.

The learning rate factor is not present in the batch method which is faster than the sequential training. The training process of both SOM algorithms can be decomposed into two phases: an ordering phase followed by a convergence phase. In the ordering phase, the learning rate parameter and neighborhood function begin with large values and then shrink slowly with time (epoch). In the convergence phase, the learning rate parameter maintains small values and the neighborhood function contains only nearest neighbors of the BMU.

B. Support vector machine classifier

SVM is a supervised binary classifier, whose goal is to divide the d-dimensional input feature space into two subspaces (one for each class) using a separating hyperplane. An important feature of SVM is related to the possibility to project the original data into a higher dimensional feature space via a kernel function $K(.,.)$, that implicitly models the classification problem into a higher dimensional space where linear separation between classes can be approximated.

Let us assume that a training set consists of N labeled samples $(x_i, y_i)_{i=1}^N$, where $x_i \in \Re^d$ are the training samples and $y_i \in \{+1, -1\}$ are the associated labels (which model classes ω_1 and ω_2). The training phase of the classifier can be formulated as an optimization problem, which by using the Lagrange optimization theory, leads to the following dual representation:

$$\begin{gathered}\max_{\alpha} \quad \left\{\sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N y_i y_j \alpha_i \alpha_j K(x_i, x_j)\right\} \\ \sum_{i=1}^N y_i \alpha_i = 0 \\ 0 \leq \alpha_i \leq C \\ i = 1, 2, ..., N \end{gathered} \tag{4}$$

where α_i are Lagrangian multipliers, and C is a regularization parameter that allows one to control the penalty assigned to errors. The decision function $f(x)$ is defined as:

$$f(x) = \sum_{x_i \in SV} \alpha_i y_i K(x_i, x) + b \tag{5}$$

where SV represents the set of support vectors. The training pattern x_i is a support vector if the corresponding α_i has a nonzero value. For a given test sample x, the sign of the discriminant function $f(x)$ defined in (5) is used to predict its class label.

In order to address multiclass problems here, we adopt the one-against-all (OAA) strategy, which involves a parallel architecture made up of n binary SVMs, one for each information class. Each SVM solves a two-class problem defined by one information class against all the others [29].

C. Proposed SOM-SVM based active learning method

The proposed method first selects the h_1 most uncertain and diverse samples by exploiting the SVM classifier and the SOM neural network. Then, by exploiting the topological property of the SOM, it incorporates the cluster assumption in the process to choose the h $(1 < h < h_1)$ most informative samples from the h_1 for labeling. Fig. 2 shows the complete block scheme of the proposed method.

The proposed technique consists of two main steps. In the first step, a SOM network is trained in an unsupervised way in order to identify the available important samples that belong to low-density regions of the feature space. This is accomplished by updating the weight vector associated to each neuron in the map so that when the network achieves convergence the weight vectors describe a mapping from the higher dimensional input feature space to a lower dimensional output/map space. In our method, at the convergence of the training phase we compute the average neighbor distance of each neuron of the map to its neighboring neurons by using their corresponding weight vectors. The average neighbor distance of neuron k, denoted as $\overline{w_k}$, is computed as follows:

$$\overline{w_k} = \frac{1}{|N_k^r|} \sum_{i \in N_k^r} \| w_k - w_i \|^2 \tag{6}$$

where N_k^r represents the set of neurons in the map that are in the r^{th} order neighbor system of the neuron k. Under the assumption that SOM preserves the topological property of the input space (see the end of this section for a discussion on this assumption), we can use the set of obtained average neighbor distance measures to identify samples that belong to low-density regions of the feature space. These samples are associated with the neurons that have larger average distance values. Accordingly we can state that due to the cluster assumption, the neurons that have higher average neighbor distance have a higher probability to map boundary samples than the neurons having lower average neighbor distance. This information is then exploited in the second step of the proposed method. It is worth noting that SOM is run only once on all (or a properly sub-sampled set) of the input samples (i.e., the pixels of the considered images). This is done before starting the iterative active learning process.

The second step of the proposed technique is aimed to select the most informative samples at each iteration of the active learning process to solve a n $(n > 1)$ class classification problem. To this end, initially we train with the available labeled samples n SVM binary classifiers (each one associated with a separate class) organized in a OAA architecture. After training, for each unlabeled sample x in the unlabeled pool U, n functional distances $f_i(x), i = 1, 2, ..., n$ are obtained which correspond to the n decision hyperplanes of the binary SVM classifiers included in the OAA architecture. Then the confidence value $s(x)$ associated with the classification reliability of each unlabeled sample $x \in U$ can be computed. It is worth noting that the confidence can be related to the uncertainty associated with the considered sample. Two alternative strategies can be used for computing the confidence [15]. The first strategy is based on the widely used marginal sampling (MS) technique [12], where the smallest distance among the n decision hyperplanes is considered to compute the confidence value $s(x)$ of each unlabeled sample $x \in U$, i.e.,

$$s(x) = \min_{i=1,2,...,n} \{|f_i(x)|\} \tag{7}$$

The second strategy is based on the multiclass label uncertainty (MCLU) technique [30]. In this technique the difference between the first and second largest distance values to the hyperplanes are considered to compute the confidence value $s(x)$ of each unlabeled sample $x \in U$ as follows:

$$s(x) = f_{r_1}(x) - f_{r_2}(x) \tag{8}$$

where

$$r_1 = \arg \max_{i=1,2,...,n} \{|f_i(x)|\}, \quad r_2 = \arg \max_{i=1,2,...,n; i \neq r_1} \{|f_i(x)|\}.$$

With both strategies the uncertainty of each unlabel sample $x \in U$ is measured according to its corresponding $s(x)$ value. The samples that have lower confidence values are considered as the most uncertain since they have the lowest correct classification confidence.

After computing the uncertainty of each unlabeled sample by using (7) or (8), we select the h_1 samples from U which have the lowest confidence values (this imposes the uncertainty criterion) and are mapped into distinct neurons of the SOM according to the results obtained in the first step (this imposes

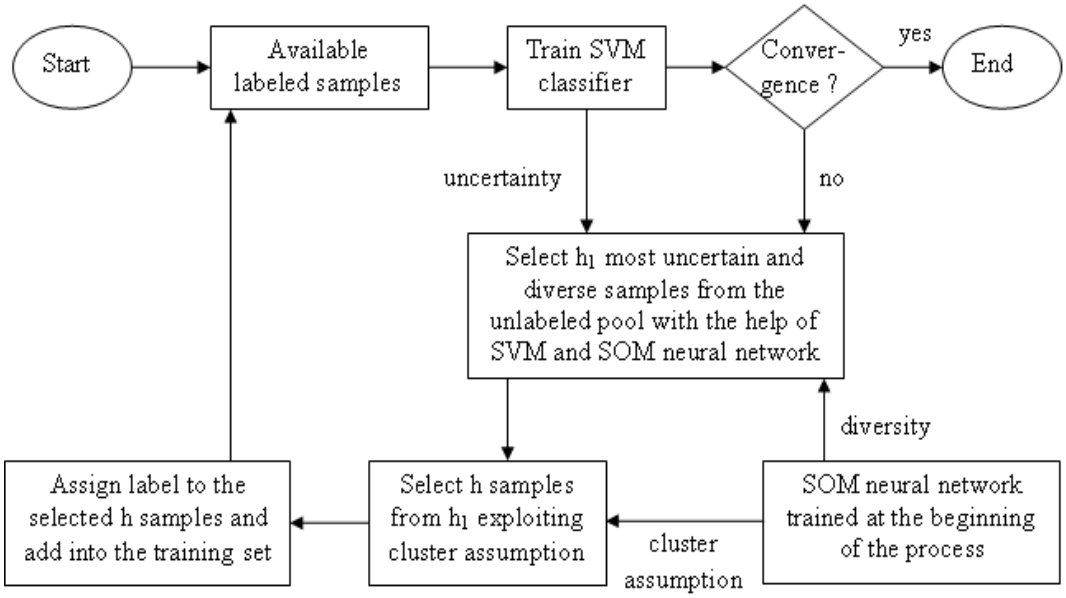


Fig. 2. Flowchart of the proposed method.

the diversity criterion). This allows us to select the h_1 most uncertain samples which are diverse from each other because similar input patterns are mapped into the same neuron. Then a batch of h $(1 < h < h_1)$ samples from the selected h_1 samples are chosen that correspond to the SOM mapping neurons having the highest average neighbor distances computed in (6). This allows us to incorporate the cluster assumption property in the selection of the h most informative samples for labeling. In other words we select samples that are both uncertain, diverse and located in low density regions of the feature space (i.e., under the cluster assumption on the boundary of decision regions). This can be particularly useful when biased initial training sets that do not model the real distribution of data close to the decision boundary are available. The process is iterated until a stop criterion (which can be related to the stability of accuracy or to its value) is satisfied. Algorithm 3 provides the details of the proposed technique.

It is worth noting that the proposed method incorporates the cluster assumption in the selection process by assuming that the SOM neural network is able to preserve the topological property of the input data. Like any other nonlinear dimensionality reduction technique, a SOM neural network does not guarantee the preservation of the topology in all kinds of problems, especially when very high dimensional feature spaces are considered [31], [32]. If for a given data set SOM fails to preserve the topology, the proposed technique may not be able to incorporate the cluster assumption criterion properly. In this case, the uncertainty and diversity criteria play the main role to select the most informative samples at each iterations of the active learning. This results in a possible increase of the number of iterations required to reach the convergence.

Algorithm 3 Proposed SOM-SVM based active learning method

1: Train the SOM neural network by using available patterns (both labeled and unlabeled).
2: Compute the average neighbor distance of each neuron using (6)
3: **repeat**
4: Train with the available initial labeled samples n binary SVM classifiers (each one associated with a specific informative class) organized in a OAA architecture.
5: Compute the confidence value of each unlabeled sample by using either (7) or (8).
6: Select the h_1 samples from U which have the lowest confidence (i.e., the lowest certainty) values and are mapped into distinct neurons of SOM (diversity criterion).
7: Select the h $(h < h_1)$ samples from the h_1 samples that correspond to the SOM mapping neurons having the highest average neighbor distances (exploitation of the cluster assumption).
8: Assign labels to the h selected samples and include them into the training set.
9: **until** the stop criterion is satisfied and the final training set is obtained.

III. Data sets description

In order to assess the effectiveness of the proposed active learning technique, four data sets with significantly different properties were used in the experiment. The first one is a toy data set which is made up of four linearly separable classes as shown in Fig. 3. It contains 1000 samples, and only 4 samples (one from each class) were chosen as initial training samples; the remaining 996 samples were in the unlabeled pool U.

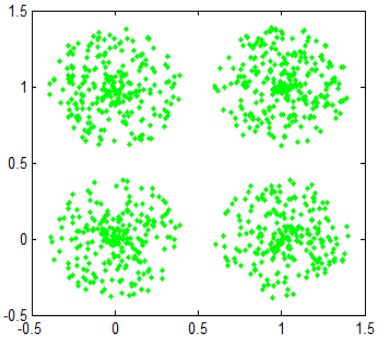


Fig. 3. Linearly separable toy data set in a two dimensional feature space.

The second data set is a simple color image as shown in Fig. 4. The image contains five different color balloons. It is used for assessing the performance of the proposed technique on a simple problem. First we manually generated some labeled samples from the image. Then these samples were randomly split into a training set T of 1553 samples and a test set TS of 1617 samples. Initially, only 10 samples (two from each class) were randomly selected from T as initial training set L, and the rest were stored in the unlabeled pool U, as shown in Table I.

Fig. 4. Color image used in the second experiment.

TABLE I
NUMBER OF SAMPLES FOR EACH CLASS IN THE INITIAL TRAINING SET(L), IN THE TEST SET(TS) AND IN THE UNLABELED POOL(U) FOR THE COLOR IMAGE DATA SET

Classes	L	TS	U
Green	2	298	306
Blue	2	308	310
Yellow	2	298	325
Pink	2	319	336
Sky blue	2	320	340
Total	**10**	**1543**	**1617**

The third data set shown in Fig. 5 is a Quickbird multispectral remote sensing image acquired on the city of Pavia (northern Italy) on June, 2002. It consists of four pan-sharpened

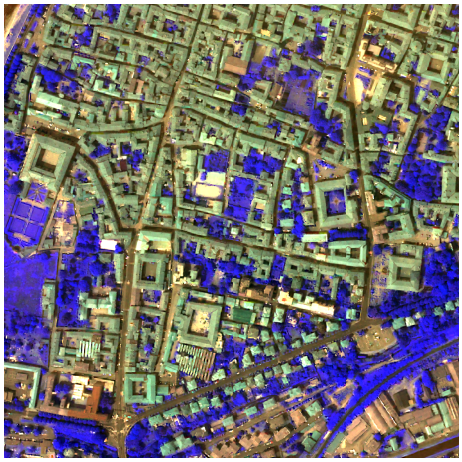


Fig. 5. Multispectral image used in our experiments.

(merging high-resolution panchromatic and lower resolution multispectral channels) bands and a panchromatic channel with a spatial resolution of 0.7m. The size of the full image is 1024×1024 pixels and there are eight classes. The reader is referred to [33] for more details on this data set. The available labeled samples were collected by photointerpretation. These samples were randomly split into a training set T of 5707 samples and a test set TS of 4502 samples. In our experiments, first only few samples were randomly selected from T as initial training set L, and the rest were stored in the unlabeled pool U. Table II shows the land-cover classes and the related number of samples used in the experiments.

TABLE II
NUMBER OF SAMPLES FOR EACH CLASS IN THE INITIAL TRAINING SET(L), IN THE TEST SET(TS) AND IN THE UNLABELED POOL(U) FOR THE MULTISPECTRAL DATA SET

Classes	L	TS	U
Water	2	215	178
Tree areas	4	391	344
Grass areas	4	321	319
Road	12	613	975
Shadow	9	666	709
Red building	29	1620	2267
Gray building	7	427	590
White building	3	249	255
Total	**70**	**4502**	**5637**

The fourth data set shown in Fig. 6 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 available labeled data were collected using land-cover maps derived from color infrared photography provided by KSC and Landsat thematic mapper imagery. The reader is referred to [34] for more details on this data set. After the elimination of noisy samples, the labeled samples were randomly split into a training set T of 5707 samples and a test set TS of 2556 samples. In our experiments, first only few samples were randomly selected from T as initial training set L, and the rest

Fig. 6. Hyperspectral image used in our experiments.

were stored in the unlabeled pool U. Table III shows the land-cover classes and the related number of samples used in the experiments.

TABLE III
NUMBER OF SAMPLES FOR EACH CLASS IN THE INITIAL TRAINING SET(L), IN THE TEST SET(TS) AND IN THE UNLABELED POOL(U) FOR THE HYPERSPECTRAL DATA SET

Classes	L	TS	U
Scrub	15	380	366
Willow swamp	5	120	116
Cabbage palm hammock	5	128	123
Cabbage palm/Oak hammock	5	125	121
Slash pine	3	80	78
Oak/Broadleaaf hammock	5	114	110
Hardwood swamp	2	52	51
Graminoid marsh	9	215	207
Spartina marsh	10	260	250
Cattaial marsh	8	188	181
Salt marsh	8	209	201
Mud flats	9	231	222
Water	18	454	536
Total	**102**	**2556**	**2463**

IV. EXPERIMENTAL RESULTS

A. Design of experiments

In our experiments we adopted an OAA architecture of SVM classifiers. Each SVM was implemented with radial basis function (RBF) kernels. 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. In all our experiments we used a number of neurons in the output layer of SOM sufficiently larger than the number of classes of the input patterns so that the samples that belong to a specific class are mapped onto a group of neighboring neurons.

The proposed technique was implemented by considering both the MS and the MCLU uncertainty criteria as described in (7) and (8), respectively. Depending on the use of the MS and the MCLU uncertainty criteria, we refer to MS-Proposed and MCLU-Proposed technique. To assess the effectiveness of the proposed method we compared it with other effective methods recently proposed in the literature and with more traditional techniques: i) the cluster assumption with histogram thresholding (CAHT) [8]; ii) the multiclass label uncertainty with enhanced cluster based diversity (MCLU-ECBD) [15]; iii) the marginal sampling by closest support vector (MS-cSV) [12]; iv) the entropy query-by-bagging (EQB) [12]; and v) the random sampling (RS). The CAHT approach, first detects a threshold for each binary SVM classifier, which identifies the low density regions of the SVMs output space. The m $(m > h)$ most uncertain samples from U are selected considering the patterns having the output scores closest to one of the selected thresholds. Then, by applying the kernel k-means clustering algorithm, the selected m samples are divided into h different clusters and the most uncertain samples from each cluster are chosen for labeling. The MCLU-ECBD first selects the m most uncertain samples from U that have minimum confidence values computed using (8). Then, by applying the kernel k-means clustering algorithm, the selected m most uncertain samples are divided into h different clusters and the sample from each cluster that is closest to the SVM decision hyperplane is chosen for labeling. The MS-cSV approach selects m most uncertain samples from U which have minimum confidence values computed using (7). Then, the h samples from the selected m patterns which do not share the same closest support vector are chosen for labeling at each iteration of the active learning process. Note that in the present experiments the value of m is fixed to $m = 3h$ for a fair comparison among the different techniques. The EQB selects the h most uncertain samples according to the maximum disagreement between a committee of classifiers. The committee is obtained by bagging: first different training sets are drawn with replacement from the original training data. Then, each training set is used to train the OAA SVM to predict the different labels for each unlabeled sample. Finally, the entropy of the distribution of the different labels associated to each sample is calculated to evaluate the disagreement among the classifiers on the unlabeled samples. In the RS approach, at each iteration a batch of h samples are randomly selected from the unlabeled pool U and included into the training set.

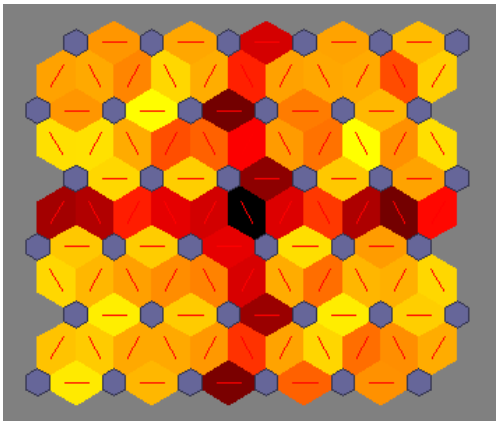
In the present experiment we used the batch algorithm to train the SOM neural network. This algorithm is implemented using Matlab (R2009b) functions. The multiclass SVM with the OAA architecture has been manually implemented by using the LIBSVM library (for Matlab interface) [35]. All the active learning algorithms presented in this paper have been implemented in Matlab.

B. Results: Toy data set (Experiment 1)

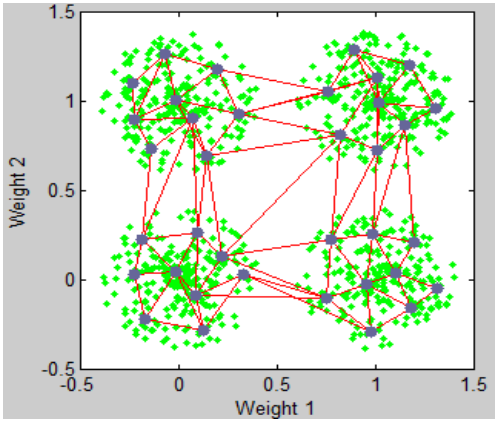
In order to understand the potential of the proposed technique and to illustrate its behavior, in the first experiment we compared the different active learning methods by using the simple toy data set described in Section III. For this data set, we constructed a SOM neural network with 6×6 neurons

arranged in a hexagonal lattice on its output layer (map). The network was trained with all the available input patterns $x \in (U \cup L)$. The network spent 500 epochs in the ordering phase and 500 epochs in the convergence phase. The initial value of the neighborhood function was set to 3, which was gradually decreased and reached 1 at the end of 500 epochs. Figs. 7 (a) and (b) show the distance between neighboring neurons in the SOM map and the position of the weight vectors in the input space, respectively, at the end of the SOM network training. From Fig. 7 (a) one can see that a group of light segments appear in the lower-left, lower-right, upper-left and upper-right regions, bounded by some darker segments. This grouping indicates that the network clustered the toy data into four groups, one for each class. These four groups can also be seen in the weight vector positions graph in Fig. 7 (b). The four light segments of Fig. 7 (a) contain four groups of tightly clustered data points where the two neighbor neurons in the same segment have the smaller distance, as indicated by lighter colors. Whereas the distances between two neighbor neurons in different segments are larger, as indicated by darker colors. This distance is also shown in the weight vector positions figure. From these two figures one can see that the patterns which are in the border regions of the input space are mapped onto the neurons which have larger average neighbor distances.

In this experiment, initially only 4 labeled samples (1 from each class) were chosen for the training and 4 additional samples were selected at each iteration of active learning. The process was iterated 5 times to have 24 samples in the training set at the end. To reduce the random effect on the results, the active learning process was repeated for 10 trials with different initial labeled samples. For a quantitative analysis, Table IV reports the classification accuracy obtained by the MS-Proposed, the MCLU-Proposed, the CAHT, the MCLU-ECBD, the MS-cSV, the EQB and the RS methods at different iterations. From the table one can see that the proposed technique obtained 100% classification accuracy after the 1st iteration (i.e., by labeling only 8 additional samples), while the CAHT technique needed at least 2 iterations (i.e., 12 samples) and the other most effective techniques (i.e., the MCLU-ECBD, and the MS-cSV) needed at least 3 iterations (i.e., 16 samples) to achieve the same accuracy. This confirms that, as the proposed technique selects the informative sample from low-density regions (border regions) of the feature space by exploiting the topological properties of SOM, it can converge fast also when the available initial training samples are poor (i.e. they are biased with respect to the representation of the structure of the classification problem). Although this is a simple example, starting from a suboptimal data set, the proposed technique, thanks to the SOM that implement the diversity criterion and the cluster assumption for selecting the informative samples, reaches the convergence decreasing of 33% the number of required new labeled samples with respect to the best literature methods used in our comparison.



(a)



(b)

Fig. 7. SOM at the convergence of the network training phase. (a) Distances between neighbor neurons arranged in the hexagonal lattice. The blue hexagons represent the neurons and the red lines connect neighboring neurons. The colors in the regions containing the red lines indicate the distances between neurons. The darker colors represent larger distances and the lighter colors represent smaller distances. (b) Positions of the weight vectors associated with the neurons in the lattice (toy data set).

TABLE IV
OVERALL CLASSIFICATION ACCURACY ($\overline{OA}$) PRODUCED BY THE DIFFERENT TECHNIQUES AT DIFFERENT ITERATIONS (TOY DATA SET)

Itr No	Training Samples	$\overline{OA}$						
		MS-Proposed	MCLU-Proposed	CAHT	MCLU-ECBD	MS-cSV	EQB	RS
0	4	98.15	98.15	98.15	98.15	98.15	98.15	98.15
1	8	98.56	98.48	98.68	98.37	99.26	97.69	98.91
2	12	100	100	99.79	99.66	99.72	98.17	99.23
3	16	100	100	100	99.95	99.99	98.94	99.24
4	20	100	100	100	100	100	99.87	99.14
5	24	100	100	100	100	100	100	99.14

C. Results: Color image data set (Experiment 2)

The second experiment was carried out to compare the performance of the proposed technique with those of other techniques by using the color image data sets described in Section III. For this data set, we constructed a SOM neural network with 6×6 neurons arranged in a hexagonal lattice on its output layer (map). The network was trained with all the available input patterns. The network spent 500 epochs in the ordering phase and 500 epochs in the convergence phase. The initial value of the neighborhood function was set to 3, which was gradually decreased and reached 1 at the end of 500 epochs. In this experiment, initially only 10 labeled samples (2 from each class) were chosen for the training

and 5 additional samples were selected at each iteration of active learning. The process was iterated 5 times to have 35 samples in the training set at the end. To reduce the random effect on the results, the active learning process was repeated for 10 trials with different initial labeled samples. For a quantitative analysis, Table V reports the classification accuracy obtained by the MS-Proposed, the MCLU-Proposed, the CAHT, the MCLU-ECBD, the MS-cSV, the EQB and the RS methods at different iterations. From the table one can see that the proposed technique obtained 100% classification accuracy after the 1st iteration (i.e., by labeling only 10 additional samples), while the other most effective techniques (i.e., the CAHT, the MCLU-ECBD, and the MS-cSV) needed at least 3 iterations (i.e., 20 samples) to achieve the same accuracy. This again confirms that, the proposed technique converge fast also when the available initial training samples are biased.

TABLE V
OVERALL CLASSIFICATION ACCURACY ($\overline{OA}$) PRODUCED BY THE DIFFERENT TECHNIQUES AT DIFFERENT ITERATIONS (COLOR DATA SET)

Itr No	Training Samples	$\overline{OA}$						
		MS-Proposed	MCLU-Proposed	CAHT	MCLU-ECBD	MS-cSV	EQB	RS
0	10	95.28	95.28	95.28	95.28	95.28	95.28	95.28
1	15	99.20	98.28	99.58	99.57	99.17	96.69	96.63
2	20	100	100	99.80	99.92	99.68	99.09	98.64
3	25	100	100	99.95	99.68	99.91	99.09	98.54
4	30	100	100	100	100	100	99.45	98.84
5	35	100	100	100	100	100	100	99.33

D. Results: Multispectral data set (Experiment 3)

The third experiment was carried out to compare the performance of the proposed technique with other techniques by using the multispectral data set described in Section III. For this data set, we constructed the SOM neural network with 25×25 neurons arranged in a hexagonal lattice on its output layer (map). The network was trained with all the available input patterns. The network spent 4000 epochs in the ordering phase and 16000 epochs in the convergence phase. The initial value of the neighborhood function was set to 20, which was gradually decreased and reached 1 at the end of 4000 epochs.

Initially only 70 labeled samples were included in the training set to train the SVM and 20 samples were selected at each iteration of active learning. The whole process was iterated 19 times resulting in 450 samples in the training set at convergence. To reduce the random effect on the results, the active learning process was repeated for 10 trials with different initial labeled samples.

Fig. 8 shows the average overall classification accuracies provided by different methods versus the number of samples included in the training set at different iterations for the multispectral data set. From this figure, one can see that the proposed active learning technique based on either MS or MCLU uncertainty criteria outperformed all other techniques. The proposed technique produced better results also compared to the CAHT, which is another cluster assumption based method. This confirms that the proposed technique properly exploits the diversity criterion and cluster assumption for selecting informative samples. For a quantitative analysis, Tables VI reports the average class accuracies (%), the mean and the standard deviation of the overall accuracy, as well as the average kappa accuracies obtained at convergence on 10 runs for the multispectral data set. From the table, one can see that the proposed technique based on the MS and MCLU criteria resulted in an overall accuracy of 87.37% and 87.40%, respectively. Among the other techniques, the highest overall accuracy (86.53%) was produced by the CAHT technique. By analyzing the class wise accuracy, the proposed technique produced highest accuracy for the larger number of classes with respect to other techniques. Moreover the standard deviation of the proposed approach is smaller than those of the other techniques. This means that, as expected from the exploitation of the cluster assumption, the proposed method is more robust to the quality of initial training samples available. This confirms the effectiveness of the proposed technique for the multispectral data set.

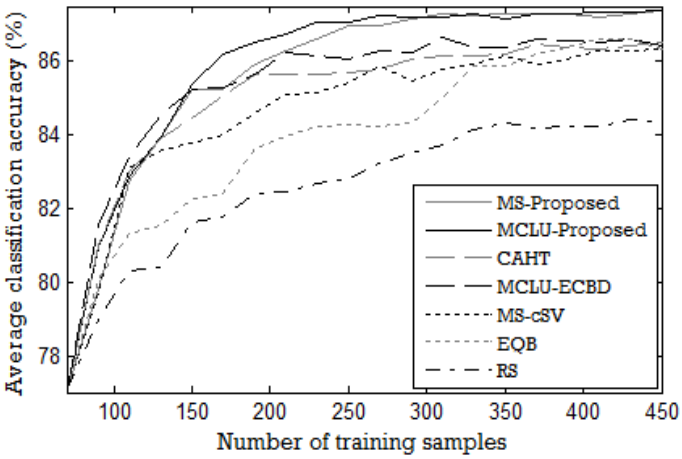


Fig. 8. Average classification accuracy over ten runs versus the number of training samples provided by the MS-Proposed, the MCLU-Proposed, the CAHT, the MCLU-ECBD, the MS-cSV, the EQB, and the RS methods (multispectral data set).

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

Methods	MS-Proposed	MCLU-Proposed	CAHT	MCLU-ECBD	MS-cSV	EQB	RS
Labeled samples	**450**	**450**	**450**	**450**	**450**	**450**	**450**
water	83.07	**83.86**	79.26	76.60	77.81	76.70	83.77
tree areas	83.22	83.60	83.48	83.02	83.53	**83.76**	79.85
grass areas	82.80	82.55	82.74	82.68	**82.90**	81.62	81.59
road	81.96	82.20	83.47	**83.69**	82.76	82.50	83.16
shadow	**87.99**	87.82	84.65	84.65	84.38	83.89	79.77
red building	**96.81**	96.48	96.73	96.46	96.53	96.70	95.56
gray building	67.56	**68.67**	65.39	66.67	65.19	68.08	57.07
white building	**87.75**	87.43	84.82	85.22	86.71	84.70	85.62
$\overline{OA}$	87.37	**87.40**	86.53	86.43	86.36	86.34	84.38
std	0.23	**0.21**	0.36	0.29	0.51	0.37	0.61
$kappa$	**0.843**	**0.843**	0.832	0.831	0.830	0.830	0.806

E. Results: Hyperspectral data set (Experiment 4)

The fourth experiment was carried out by using the hyperspectral data set described in Section III. For this data

set, the architecture of the SOM neural network and its learning parameters are defined exactly same as described in the experiment 3.

Initially only 102 labeled samples were included in the training set to train the SVM and 20 samples were selected at each iteration of active learning. The whole process was iterated 20 times resulting in 502 samples in the training set at convergence.

Fig. 9 shows the average overall classification accuracies provided by different methods versus the number of samples included in the training set at different iterations for the hyperspectral data set. From this figure, one can see that the MCLU-Proposed approach converged faster (in terms of number of labeled samples) and produced better results than the all other techniques. For a quantitative analysis, Tables VII reports the average class accuracies, mean and standard deviation of the overall accuracies, as well as the average kappa accuracies obtained at convergence on 10 runs for the hyperspectral data set. From the table, one can see that the MCUL-Proposed technique resulted in an overall classification accuracy of 95.18%. Among the other techniques, the best overall accuracy (94.91%) was produced by the MCLU-ECBD technique. An analysis of the class wise accuracies points out that the proposed technique yield the highest accuracy for a larger number of classes than the other methods. Moreover, also in this case the standard deviation of the MCLU-Proposed approach is smaller than those of the other techniques, thus pointing out its high robustness to initial training conditions. The MS-Proposed technique produced similar results as produced by the CAHT and the MS-cSV techniques. These methods failed to find optimal solution for the highly mixed class "cabbage palm/hammock", which is a critical class because highly overlapped in the feature space to the others. This is because of the uncertainty criterion (MS) used by these techniques is not able to select proper samples from that class.

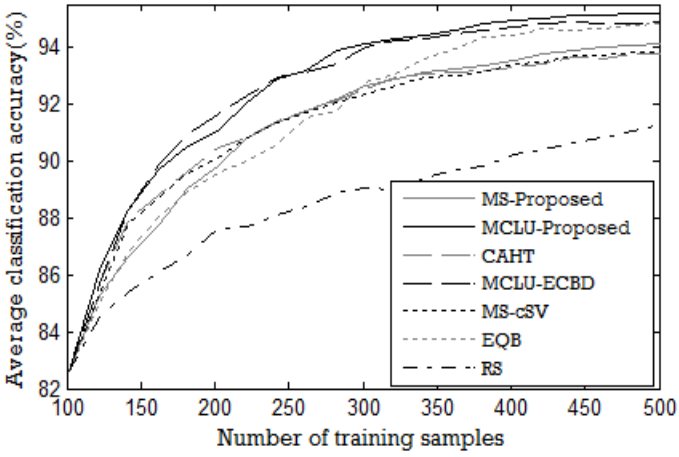


Fig. 9. Average classification accuracy over ten runs versus the number of training samples provided by the MS-Proposed, the MCLU-Proposed, the CAHT, the MCLU-ECBD, the MS-cSV, the EQB, and the RS methods (hyperspectral data set).

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

Methods	MS-Proposed	MCLU-Proposed	CAHT	MCLU-ECBD	MS-cSV	EQB	RS
Labeled samples	502	502	502	502	502	502	502
scrup	97.37	96.97	**97.39**	96.97	97.24	96.71	95.45
willow swa.	94.00	**95.25**	93.17	92.16	92.33	90.58	89.50
cabb. pl. ham.	92.97	90.86	92.50	93.67	93.67	**94.14**	90.47
cabb. pl./oak ham.	56.40	**77.60**	62.64	72.56	58.00	70.40	52.88
slash pn.	75.25	78.13	68.62	77.62	73.12	**78.62**	69.37
oak/broad. ham.	**79.39**	78.60	74.38	78.68	77.81	78.86	63.33
hardwood swa.	87.31	88.08	83.84	88.46	88.65	**92.31**	82.88
graminoid mr.	93.95	94.42	93.53	**94.47**	93.44	**94.47**	91.81
spartina mr.	99.54	**99.50**	99.34	99.35	99.42	**99.50**	98.77
cattaial mr.	99.79	**99.95**	99.68	99.89	99.41	**99.95**	98.30
salt mr.	**100**	99.95	99.90	**100**	99.71	99.95	97.85
mud flats	97.92	**98.00**	97.53	98.05	97.14	97.97	92.59
water	99.87	**99.98**	99.96	99.93	99.89	99.89	99.93
$\overline{OA}$	94.11	**95.18**	93.76	94.91	93.84	94.83	91.31
std	0.41	**0.18**	0.25	0.23	0.27	0.31	.56
$kappa$	0.934	**0.946**	0.930	0.943	0.931	0.942	0.903

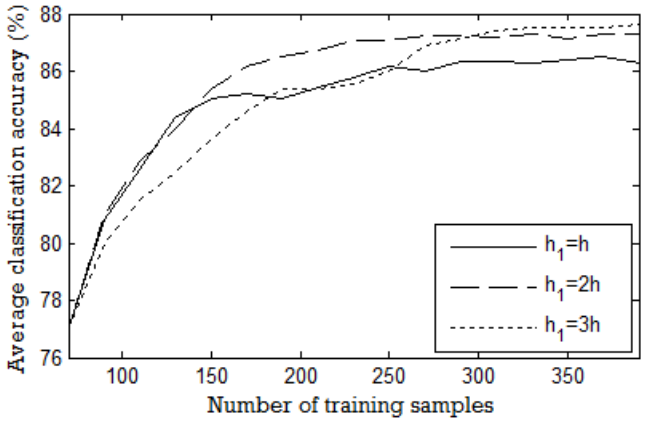
F. Results: Sensitivity analysis (Experiment 5)

In the fifth experiment, we analyzed the performance of the proposed technique by varying the value of h_1, i.e., the number of most uncertain and diverse samples chosen before applying the cluster assumption to select the h informative samples to be labeled. Figs. 10(a) and (b) show the classification results obtained by the MCLU-Proposed technique by considering different values of h_1 for the multispectral and the hyperspectral data sets, respectively. From these figures one can see that, when the proposed technique selected h samples without involving the cluster assumption (i.e., $h_1 = h$), it achieved similar results as those produced by the best uncertainty and diversity based active learning approaches. On the contrary, when the proposed technique exploited the cluster assumption in the selection of the h informative samples from the selected h_1 most uncertain and diverse samples, it converged with increased accuracies for both the considered data sets. This confirms the importance in using the information on the low density regions of the feature space in the query function. From the figures one can also see that when the value of h_1 increases, the proposed technique needs a higher number of labeled samples for converging, due to the inclusion of some samples which are not uncertain enough. In our experiment, we found that $h_1 = 2h$ is a suitable choice for the sample selection process of the proposed technique. Note that the same analysis is also valid for the MS-Proposed approach.

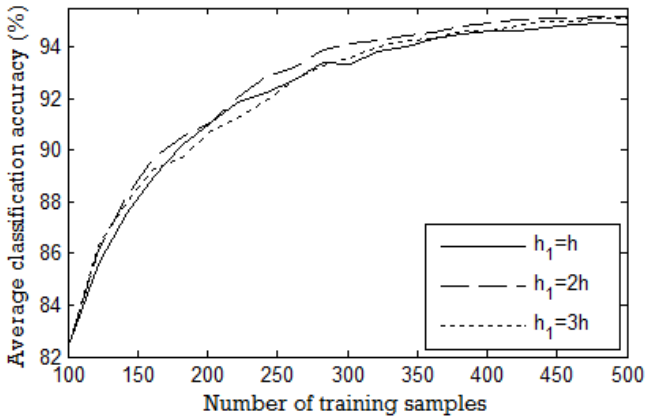
Finally, we carried out different experiments for assessing the stability of the proposed technique both by increasing the number of neurons on the map and by varying the initial value of the neighborhood function for the SOM neural network. The results of all these experiments pointed out the almost insensitivity of the proposed algorithm to these parameters.

V. DISCUSSION AND CONCLUSION

In this paper, a novel batch mode active learning technique for solving remote sensing image classification problems has been proposed. The query function of the proposed technique is modeled by incorporating uncertainty, diversity and cluster



(a)



(b)

Fig. 10. Average classification accuracy provided by the MCLU-Proposed technique with different value of h_1 for (a) the multispectral and (b) the hyperspectral data sets.

assumption criteria. The uncertainty criterion is incorporated by exploiting the properties of SVM classifiers according to either the marginal sampling or the multiclass label uncertainty techniques. The diversity and cluster assumption criteria are incorporated by exploiting the properties of SOM neural networks. This is done by selecting uncertain samples that are mapped to neurons that both are different and have largest average neighbor distance. As the proposed technique selects informative samples from low-density regions of the feature space by exploiting the topological properties of SOM neural network, it has fast converges also when the available initial training samples are poor (i.e., the training set is bias). To assess the effectiveness of the proposed technique we compared it with others batch mode active learning techniques existing in the remote sensing literature using a toy data set and color, multispectral, and hyperspectral images. The results of this comparison pointed out that the proposed method always provided higher accuracies with improved stability with respect to those achieved by some of the most effective active learning techniques presented in the literature.

The computational time taken by the proposed technique for selecting informative samples at each iteration of the active learning process is similar to the one of the simple MS based approach. However, the proposed method requires some additional time to train the SOM network before the iterative active learning is started. This time can be reduced by regularly sub-sampling the pixels of the considered image to be used as input to SOM network.

As future developments of this work, we plan to extend the experimental comparison to other prototype based active learning methods existing in the literature. Moreover, we plan to incorporate the spatial information to the present active learning framework for improving the classification results [36].

ACKNOWLEDGMENTS

This work was carried out in the framework of the India-Trento Program for Advanced Research.

REFERENCES

[1] 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.

[2] C. Campbell, N. Cristianini, and A. Smola, "Query learning with large margin classifiers," in *Proc. 17th ICML*, 2000, pp. 111–118.

[3] 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.

[4] P. Mitra, C. A. Murthy, and S. K. Pal, "A probabilistic active support vector learning algorithm," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, pp. 413–418, 2004.

[5] M. Li and I. K. Sethi, "Confidence-based active learning," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1251–1261, 2006.

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

[7] R. Liu, Y. Wang, T. Baba, D. Masumoto, and S. Nagata, "SVM-based active feedback in image retrieval using clustering and unlabeled data," *Pattern Recognition*, vol. 41, pp. 2645–2655, 2008.

[8] 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.

[9] A. J. Joshi, F. Porikli, and N. P. Papanikolopoulos, "Scalable active learning for multiclass image classification," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2259–2273, 2012.

[10] 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.

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

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

[13] S. Patra and L. Bruzzone, "A fast cluster-assumption based active learning technique for classification of remote sensing images," *IEEE Trans. Geoscience and Remote Sensing*, vol. 49, no. 5, pp. 1617–1626, 2011.

[14] 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.

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

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

[17] ——, "A novel som-based active learning technique for classification of remote sensing images with SVM," in *Proc. IGARSS*, 2012, pp. 6879–6882.

[18] B. Demir, F. Bovolo, and L. Bruzzone, "Detection of land-cover transitions in multitemporal remote sensing images with active learning based compound classification," *IEEE Trans. Geoscience and Remote Sensing*, vol. 50, no. 5, pp. 1930–1941, 2012.

[19] 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. Geoscience and Remote Sensing*, vol. 52, no. 2, pp. 1272–1284, 2014.
[20] ——, "An effective strategy to reduce the labeling cost in the definition of training sets by active learning," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 1, pp. 79–83, 2014.
[21] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 606–617, 2011.
[22] T. Kohonen, *Self-Organizing Maps*. 2nd edn. Springer-Verlag, Boston, 1997.
[23] V. N. Vapnik, *The Nature of Statistical Learning Theory*. 2nd ed., New York: Springer, 2001.
[24] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annual Workshop Computational Learning Theory*, 1992, pp. 144–152.
[25] P. Rigollet, "Generalization error bounds in semi-supervised classification under the cluster assumption," *J. Machine Learning Research*, vol. 8, pp. 1369–1392, 2007.
[26] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
[27] R. D. Lawrence, G. S. Almasi, and H. E. Rushmeier, "A scalable parallel algorithm for self-organizing maps with applications to sparse data mining problems," *Data Mining and Knowledge Discovery*, vol. 3, pp. 171–195, 1999.
[28] S. Haykin, *Neural networks - A comprehensive foundation*. Pearson Education, Singapore, 2003.
[29] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.
[30] A. Vlachos, "A stopping criterion for active learning," *Comput. Speech Lang.*, vol. 22, no. 3, pp. 295–312, 2008.
[31] T. Villmann, R. Der, M. Herrmann, and T. M. Martinetz, "Topology preservation in self-organizing feature maps: exact definition and measurement," *IEEE Trans. Neural Networks*, vol. 8, no. 2, pp. 256–266, 1997.
[32] J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*. Springer, Information Science and Statistics series, 2007.
[33] L. Bruzzone and L. Carlin, "A multilavel context-based system for classification of very high spatial resolution images," *IEEE Trans. Geoscience and Remote Sensing*, vol. 44, no. 9, pp. 2587–2600, 2006.
[34] J. Ham, Y. Chan, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, 2005.
[35] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machine*, 2001, software available at http://csie.ntu.edu.tw/\~cjlin/libsvm.
[36] I. Dopido, J. Li, P. R. Marpu, A. J. Plaza, J. M. Bioucas Dias, and J. A. Benediktsson, "Semisupervised self-learning for hyperspectral image classification," *IEEE Trans. Geoscience and Remote Sensing*, vol. 51, no. 7, pp. 4032–4044, 2013.