This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

A Fast Cluster-Assumption Based Active-Learning Technique for Classification of Remote Sensing Images

Swarnajyoti Patra, and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—In this paper, we propose a simple, fast, and reliable active-learning technique for solving remote sensing image classification problems with support vector machine (SVM) classifiers. The main property of the proposed technique consists in its robustness to biased (poor) initial training sets. The presented method considers the 1-D output space of the classifier to identify the most uncertain samples whose labeling and inclusion in the training set involve a high probability to improve the classification results. A simple histogram-thresholding algorithm is used to find out the low-density (i.e., under the cluster assumption, the most uncertain) region in the 1-D SVM output space. To assess the effectiveness of the proposed method, we compared it with other active-learning techniques proposed in the remote sensing literature using multispectral and hyperspectral data. Experimental results confirmed that the proposed technique provided the best tradeoff among robustness to biased (poor) initial training samples, computational complexity, classification accuracy, and the number of new labeled samples necessary to reach convergence.**

***Index Terms*—Active learning, cluster assumption, entropy, hyperspectral imagery, multispectral imagery, query function, remote sensing, support vector machines (SVMs).**

I. Introduction

In REMOTE sensing literature, several supervised methods have been proposed for the classification of multispectral and hyperspectral data. All these methods require labeled samples to train the classifier, and the classification results rely on the quality of the labeled samples used for learning. Therefore, the training samples should fully represent the statistics of all the land-cover classes. However, the collection of labeled samples is time consuming and costly, and the available training samples are often not enough for an adequate learning of the classifier. Moreover, redundant samples are often included in the training set, thus slowing down the training step of the classifier without adding information. In order to reduce the cost of labeling, the training set should be kept as small as possible, avoiding redundant samples and including patterns which contain the largest amount of information and thus can optimize the performance of the model. Two popular machine learning approaches for dealing with this problem are semisupervised learning [1], [2] and active learning [3]. Semisupervised algorithms incorporate the unlabeled data into the classifier training phase to obtain more precise decision boundaries. In active learning, the learning process repeatedly queries unlabeled samples to select the most informative samples and updates the training set on the basis of a supervisor who attributes the labels to the selected unlabeled samples. In this way, unnecessary and redundant samples are not included in the training set, thus greatly reducing both the labeling and computational costs. This is particularly important for remote sensing images that may have highly redundant pixels.

In this paper, we propose a fast active-learning technique for solving multiclass remote sensing image classification problems with support vector machine (SVM) classifiers. The proposed technique is based on the cluster assumption, which is extensively used in the semisupervised classification but is seldom considered in the definition of query functions in active learning.[1] In contrast to the semisupervised learning, the query function of the proposed approach selects samples from the unlabeled pool, which have maximum ambiguity to belong to each class and are located in low-density regions of the kernel space. Initially, the SVM classifier is trained with a small number of labeled samples. After training, a histogram is constructed in the 1-D output space of the classifier by considering the output scores of the unlabeled samples in $[-1, +1]$. Since the classifier ranks each sample from the most likely members to the most unlikely members of a class, the samples whose output scores fall in the valley region of the histogram (low-density region of the kernel space) are the most uncertain/ambiguous. Thus, we can work in the 1-D output space of the classifier to identify the uncertain samples by finding the minimum value on the histogram which is associated with this uncertain region. Here, Kapur's entropy-based histogram-thresholding technique [5] is applied to detect this value. Then, a batch of samples are selected from the unlabeled pool, whose output scores are closest to the selected value. In this way, we transform the original feature space into a 1-D space [6], thus simplifying the query functions that looks for a threshold in the output space. Furthermore, since the proposed technique selects the unlabeled samples from low-density regions in the kernel space, it is not strongly affected by the initial training samples and

Manuscript received April 21, 2010; revised June 25, 2010 and August 23, 2010; accepted September 19, 2010.

The authors are with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: patra@disi.unitn.it; lorenzo.bruzzone@ing.unitn.it).

Digital Object Identifier 10.1109/TGRS.2010.2083673

[1]Cluster assumption: Two points are likely to have the same class label if there is a path connecting them, passing through high-density regions only. In other words, the decision boundary has to lie in the low-density regions of the feature space [4].

0196-2892/$26.00 © 2010 IEEE

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

by the previous training results, thus allowing relatively fast convergence also when starting with biased (poor) initial training samples. This is an important advantage in remote sensing problems where biased training sets are often available in real applications. Compared with existing methods, the proposed technique is robust to the choice of initial training samples and efficient in terms of computational complexity. Furthermore, it is able to solve multiclass classification problem with an accuracy that is comparable to those of standard active-learning techniques.

The proposed method is compared with several other active-learning methods existing in the literature by using three data sets. The first data set is an illustrative toy example. The second data set is a hyperspectral image acquired on the forest of Paneveggio, close to the city of Trento (northern Italy). The third data set is a Quickbird multispectral image acquired on the city of Pavia (northern Italy). Experimental results show the effectiveness of the proposed approach.

The rest of this paper is organized as follows. Section II describes the active-learning process and briefly surveys existing active-learning methods. The proposed cluster-assumption based active-learning approach is presented in Section III. Section IV provides the description of the three data sets used for experiments. Section V presents different experimental results obtained on the considered data sets. Finally, Section VI draws the conclusion of this work.

II. Active Learning

A general active learner can be modeled as a quintuple (G, Q, S, L, and U) [6]. G is a classifier, which is trained on the labeled samples in the training set L. Q is a query function used to select the most informative samples from an unlabeled sample pool U. S is a supervisor who can assign the true class label to the selected samples from U. Initially, the training set L has few labeled samples to train the classifier G. After that, the query function Q is used to select a set of samples from the unlabeled pool U, and the supervisor S assigns a class label to each of them. Then, these new labeled samples are included into L, and the classifier G is retrained using the updated training set. The closed loop of querying and retraining continues for some predefined iterations or until a stop criterion is satisfied. *Algorithm 1* gives a description of a general active-learning process.

Algorithm 1: Active-learning process

Step 1: Train the classifier G with the training set L (which initially has few labeled samples).

Repeat

- **Step 2:** Select a set of samples from the unlabeled pool U using the query function Q.
- **Step 3:** Assign a class label to each of the queried samples by a supervisor S.
- **Step 4:** Add the new labeled samples to the training set L.
- **Step 5:** Retrain the classifier G.

Until the stop criterion is satisfied

The query function is fundamental in the active-learning process. Several methods have been proposed in the machine learning literature which differ only in their query functions. A probabilistic approach is presented in [7], which is based on the estimation of posterior probability density function. For two-class cases, the uncertain samples are identified by choosing the patterns whose class membership probability is closest to 0.5. The query function proposed in [8] is designed to minimize future errors. This approach is applied to two regression problems where an optimal solution for minimizing future error rates can be obtained in closed form. Unfortunately, for most classifiers, it is not possible to calculate the expected error rate without specific statistical models. In [9], Fukumizu has proposed a statistical active-learning approach to train multilayer perceptron for performing regression.

Another class of active-learning methods is based on query by committee [10]–[12], wherein the sample that has the highest disagreement among the committee of classifiers is chosen for the labeling. The algorithm theoretically guarantees the reduction in prediction error with the number of iterations. Variations of the query-by-committee algorithm, such as query-by-bagging and query-by-boosting algorithms, have been presented in [13] and [14].

An interesting category of active-learning methods is based on SVMs, which have gained significant success in many real-world applications, including remote sensing [15]–[20]. The SVM classifiers [21]–[23] are particularly suitable for active learning due to their intrinsic high-generalization capabilities and because their classification rule is characterized by a small set of support vectors that can be easily updated over successive learning iterations [19]. One of the most popular and effective query heuristics for SVM active learning is to select the data point closest to the current-separating hyperplane, which is also referred to as marginal sampling (MS) [15]. An active-learning strategy based on version space splitting is presented in [17]. The algorithm attempts to select the points that split the current version space into two halves having equal volumes at each step. Three heuristics for approximating the aforementioned criterion are described; the simplest among them selects the point closest to the current hyperplane [15]. In [16], a greedy optimal strategy based on SVM is presented for active learning.

It is important to note that most of the aforementioned methods consider only one sample at each iteration. However, in many problems it is necessary to speed up the learning process by selecting batches of more than one sample at each iteration. In [24], Mitra *et al.* have presented a probabilistic active-learning approach, wherein query samples are selected according to both the distance from the current-separating hyperplane and a confidence factor estimated from a set of test samples using the nearest neighbor technique. In [25], Roy and McCallum presented an active-learning approach that queried a batch of samples at each step by estimating the future error rate of each sample using two different methods. In [6], an approach that estimates the uncertainty level of each sample according to the output score of a classifier and selects only those samples whose outputs are within the uncertainty range is proposed. In [26], Brinker has presented a method that

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

selects a batch of samples by incorporating a diversity measure that considers the angles between the induced classification hyperplanes. Clustering-based active-learning approaches that query batch of samples are discussed in [20] and [27].

Active-learning methods have been scarcely considered in remote sensing image classification. Some preliminary works about the use of active learning for remote sensing image classification problems can be found in [19], [28], and [29]. In [19], an MS-based approach that selects the most uncertain sample for each binary SVM by using one-against-all (OAA) architecture to solve n-class ($n > 1$) problems is presented. In [28], two batch-mode active-learning techniques for multiclass remote sensing image classification problems are proposed. The first technique is MS by closest support vector (MS-cSV), which considers the smallest distance of the unlabeled samples to the n hyperplanes (associated to the n binary SVMs in an OAA architecture) as the uncertainty value. At each iteration, the most uncertain unlabeled samples, which do not share the closest support vector, are added to the training set. The second technique, which is called entropy query by bagging (EQB), is based on the selection of unlabeled 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 architecture 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. The samples with maximum entropy (i.e., those with maximum disagreement among the classifiers) are added to the current training set. In [29], an active-learning technique is presented, which selects the unlabeled sample that maximizes the information gain between the *a posteriori* probability distribution estimated from the current training set and the training set obtained by including that sample into it. The information gain is measured by the Kullback–Leibler divergence.

III. Proposed Method

Here, we present a fast active-learning technique based on the cluster assumption for solving remote sensing image classification problems with SVM classifier. The choice to use SVMs depends on their solid mathematical and statistical foundation and excellent performance in many real-world applications. Before describing the proposed technique, in the next section we briefly recall basic concepts related to SVM classification and introduce notation. For details, we refer to [22] and [30].

A. Support Vector Machine

Let us assume that a training set consists of N labeled samples $(x_i, y_i)_{i=1}^{N}$, where $x_i \in \Re^d$ denotes the training samples and $y_i \in \{+1, -1\}$ denotes the associated labels (which model classes ω_1 and ω_2). The goal of a binary SVM is to find out a hyperplane that separates the d-dimensional feature space into two subspaces (one for each class).

An interesting feature of SVMs is related to the possibility to project the original data into a higher dimensional feature space via a kernel function $K(.,.)$, which satisfies the Mercer conditions [30]. The training phase of the classifier can be formulated as an optimization problem by using the Lagrange optimization theory, which leads to the following dual representation:

$$Maximize: \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j K(x_i x_j)$$

$$Subject\ to: \quad \sum_{i=1}^{N} y_i \alpha_i = 0$$

$$0 \leq \alpha_i \leq C$$

$$i = 1, 2, \ldots, N$$

where α_i denotes the Lagrangian multipliers, and C is a regularization parameter that allows one to control the penalty assigned to errors. The solution to the SVM learning problem is a global maximum of a convex function. The decision function $f(x)$ is defined as

$$f(x) = \sum_{x_i \in SV} \alpha_i y_i K(x_i, x) + b \tag{1}$$

where SV represents the set of support vectors. The training pattern x_i is a support vector if the corresponding α_1 has a nonzero value. For a given test sample x, the sign of the discriminant function $f(x)$ defined in (1) is used to predict its class label.

B. Cluster-Assumption Based Active Learning

1) *Basic concept:* In the proposed approach, we estimate the uncertainty of each sample according to the output score of the SVM classifier [6]. Initially, the classifier is trained with the few available (and possibly biased) labeled samples. After training, a histogram is constructed in the 1-D output space of the classifier by considering the output scores of the samples in $[-1, +1]$. In the histogram, the region of interest is quantized into N mutually exclusive intervals called bins. We assume that all bins have equal widths (uniform quantization). The probability to have the output in a given bin is given by the number of samples whose output scores fall in that bin divided by the total number of samples in the histogram (i.e., the samples given as input to the classifier). Since the classifier ranks samples from the most likely members to the most unlikely members of a class, according to the cluster assumption (the decision boundary has to lie in low-density regions [4] of the kernel space), the samples whose output scores fall in the valley region of the histogram are the most uncertain. Thus, we can work in the 1-D output space of the classifier to identify the uncertain samples by finding a threshold on the histogram which is passing through this valley region, as shown in Fig. 1. This avoids the complexity of the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

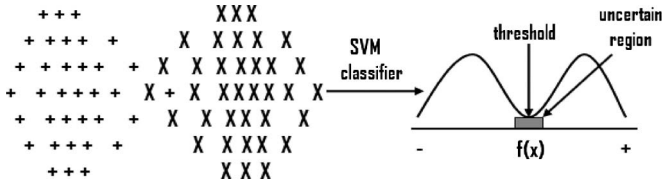

Fig. 1. Transformation of the original feature space into the 1-D classifier output space.

design of the query function in the original feature space which may have complicated decision regions and, thus, boundaries. To detect a proper threshold on the histogram, any thresholding technique existing in the pattern recognition literature can be used [31]. In this paper, we used Kapur's entropy-based histogram-thresholding technique [5], which is briefly described as follows.

2) *Entropy-based histogram thresholding:* In Kapur's method, an optimal threshold is determined based on the concept of entropy, as described in [5]. Let ω_1 and ω_2 be two classes and H be the histogram of N bins generated by considering the output scores of the SVM classifier. Let $p_i (i = 1, \ldots, N)$ be the probability of the ith bin. Assuming a threshold $t, t \in \{1, 2, \ldots, N\}$, the entropies of the classes ω_1 and ω_2 (denoted as $E_{\omega_1}(t)$ and $E_{\omega_2}(t)$, respectively) are computed as follows:

$$E_{\omega_1}(t) = -\sum_{i=0}^{t} \frac{p_i}{P_{\omega_1}(t)} \log_2 \left(\frac{p_i}{P_{\omega_1}(t)} \right)$$

$$E_{\omega_2}(t) = -\sum_{i=t+1}^{N} \frac{p_i}{P_{\omega_2}(t)} \log_2 \left(\frac{p_i}{P_{\omega_2}(t)} \right) \quad (2)$$

where $P_{\omega_1}(t) = \sum_{i=0}^{t} p_i$ and $P_{\omega_2}(t) = 1 - P_{\omega_1}(t)$. To select a threshold on the histogram that separates the two classes ω_1 and ω_2 in the output space (i.e., that passes through the valley region of the histogram), we compute the entropy of classes ω_1 and ω_2 by assuming all possible values of the threshold t. Then, the optimal threshold t_0 is selected by maximizing the total entropy $E_{\omega_1}(t) + E_{\omega_2}(t)$, i.e.,

$$t_0 = \arg \max_{t \in (1,2,\ldots,N)} \{E_{\omega_1}(t) + E_{\omega_2}(t)\}. \quad (3)$$

3) *Multiclass active-learning algorithm:* As stated before, SVMs are binary classifiers. However, several strategies have been proposed to address multiclass problems with SVMs. In order to define a multiclass architecture based on different binary classifiers, the general approach consists of defining an ensemble of binary classifiers and combining them according to a given decision rule [23]. The design of the ensemble of binary classifiers involves the definition of a set of two-class problems, each one modeled with two groups of classes. The two most commonly adopted strategies for designing the ensemble are the one against all (OAA) and the one against one (OAO) [23]. In this paper, we adopt the OAA strategy, which is based on a parallel architecture made up of n SVMs, one for each information class. Each SVM solves a two-class problem defined by one information class against all the others.

In the proposed technique, we consider each binary SVM classifier and separately select q (with q greater or equal to one) uncertain samples on the basis of the proposed query function. The q selected samples are those that, in U, have output scores closest to the detected threshold of the histogram generated by the output of the classifier. The threshold for each binary SVM is automatically detected by applying the entropy-based histogram-thresholding method described earlier. In greater detail, if we have n classes, n binary SVMs are initially trained with the current training set, and then, the functional distance $f_i(x)(i = 1, \ldots, n)$ is calculated for each binary SVM and for all the unlabeled samples $x \in U$. Then, the related histogram H_i is generated by considering the output score value in $[-1, +1]$. Thus, each binary SVM classifier generates a separate histogram considering its output score values. Then, a threshold t_i is selected for each histogram H_i by applying the entropy-based technique. Considering the ith binary SVM classifier, the q uncertain samples whose output score is the closest to the threshold t_i are selected. If, for a given classifier, there are no patterns whose output scores are in $[-1, +1]$, then the process of extraction of unlabeled pattern is stopped for that classifier. Thus, a total of $h \leq q \times n$ samples are chosen from the n binary SVM classifiers by considering only their uncertainty measure (h is lower than $q \times n$ if at least one sample is selected by more than one binary SVM or if there is at least one binary SVM which selects less than q samples). It is worth noting that a possible alternative would be to analyze also the diversity of samples for selecting the q patterns according to literature methods [26], [28]. Nonetheless, this would increase the computational time of the algorithm. Since our main goal is to have a fast technique, we prefer to avoid the use of this additional computation. The process is iterated until a stop criterion (which can be related to the stability of accuracy or to its value) is satisfied. *Algorithm 2* describes the details of the proposed technique. It is worth noting that, in the multiclass case, when the OAA architecture is used, each binary SVM has potentially different low-density regions in the input (and then in the kernel) feature space, which are associated with the boundaries between the different pair of data classes. Nonetheless, this is not a problem with the proposed technique because the use of initial training samples (even if biased) locates the SVM hyperplane close the decision boundary between the considered class and all the others and thus close to the corresponding low-density region. This is then implicitly mapped into the SVM output space that is considered by the proposed algorithm for selecting samples.

As a final remark, we point out that the proposed technique is conceptually significantly different from MS [15]. In the MS, the samples closest to the discriminant hyperplane defined at the considered iteration are selected, whereas the proposed technique looks for low-density regions in the output space of the SVM for selecting uncertain samples. The proposed approach is similar to MS only if the SVM hyperplane at the initial iteration is in the low-density region of the SVM output space.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

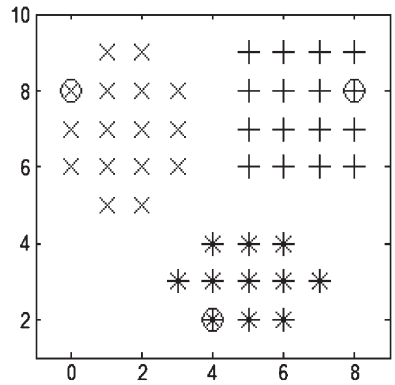

Fig. 2. Toy data set: The points represented with circles denote the initial training samples.

Algorithm 2: Proposed fast cluster-assumption based active-learning technique

Step 1: Train n binary SVMs by using a small number of labeled samples. Let $f_i(.)$ be the decision function of the ith binary SVM classifier.

Repeat

Step 2: $h = 0$.

For $i = 1$ **to** n

If $Cardinality(|f_i(x)| \leq 1) > q$

Step 3: For the ith binary SVM classifier, generate the corresponding histogram H_i by considering the output score of the unlabeled samples $x \in U$, whose output value $f_i(x) \in [-1, +1]$.

Step 4: Detect the threshold t_i from the histogram H_i by using the entropy-based histogram-thresholding technique.

Step 5: For the ith binary SVM classifier, select the q samples from the pool U, whose output scores are closest to the threshold t_i.

Step 6: $h = h + q$.

Else

Step 7: For the ith binary SVM classifier, select the samples from the pool U, whose output scores $f_i(x) \in [-1, +1]$.

Step 8: $h = h + Cardinality(|f_i(x)| \leq 1)$.

End if

End for

Step 9: Assign true labels to the h selected samples, and update the training set.

Step 10: Retrain the n binary SVMs by using the updated training set.

Until the stop criterion is satisfied.

IV. Data Set Description

Three data sets were used in the experiments. The first one is a toy data set which is made up of three linearly separable classes, as shown in Fig. 2. It contains 43 samples, and only three samples (one from each class) are chosen as initial training samples; the remaining 40 samples are in the unlabeled pool U.

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

Classes	L	TS	U
Picea Abies	39	1135	1515
Larix Decidua	13	308	520
Pinus Mugo	6	160	234
Alnus Viridis	3	70	122
No Forest	40	1000	1560
Total	101	2673	3951

TABLE II
NUMBER OF SAMPLES OF EACH CLASS IN THE INITIAL TRAINING SET (L), IN THE TEST SET (TS) AND IN THE POOL (U) FOR THE PAVIA 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 second data set is made up of a hyperspectral image acquired on the forest of Paneveggio, near the city of Trento (northern Italy) in July 2008. It consists of 12 partially overlapping images acquired by an AISA Eagle sensor in 126 bands ranging from 400 to 990 m with a spectral resolution of about 4.6 nm and a spatial resolution of 1 m. The size of the full image is 2199 $\times$ 2965 pixels. The available labeled samples were collected by ground survey. These samples were randomly split into a training set T of 4052 samples and a test set TS (to compute the classification accuracy of the algorithms) of 2673 samples. First, only few samples (2.5%) were randomly selected from T as the initial training set L, and the rest were considered as unlabeled samples stored in the unlabeled pool U. Table I shows the land-cover classes and the related number of samples used in the experiments.

The third data set is a Quickbird multispectral image acquired on the city of Pavia (northern Italy) in June 2002. It has four pan-sharpened multispectral bands and a panchromatic channel with a spatial resolution of 0.7 m. The image size is 1024 $\times$ 1024 pixels. 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. First, only few samples (1.25%) were randomly selected from T as the 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.

V. Experimental Results

A. Design of Experiments

In our experiments, we adopted an SVM classifier with radial basis kernel functions. The SVM parameters $\{\sigma, C\}$ were derived by applying the cross-validation technique. C is

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

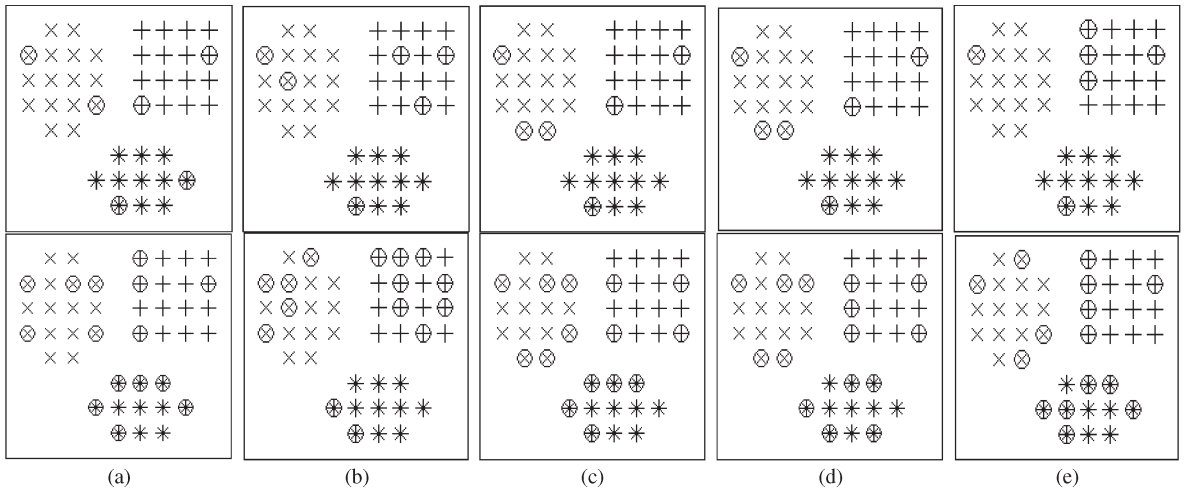

Fig. 3. Toy example showing a linear classification problem with three classes. The samples represented with circles denote the training samples selected by the (a) proposed, (b) RS, (c) MS, (d) MS-cSV, and (e) EQB methods after the (upper part of the figure) first and (lower part of the figure) fourth iterations.

a parameter controlling the tradeoff between model complexity and training error, while σ is the spread of the Gaussian kernel. The cross-validation procedure aims at selecting the best values for the parameters of the initial SVM.

To assess the effectiveness of the proposed approach, we compared it with four other methods: 1) simple random sampling (RS); 2) MS; 3) MS-cSV; and 4) EQB. The last two methods have been recently presented in remote sensing literature [28]. 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. MS is the most popular approach in the active-learning literature. In this approach, at each iteration, only the single pattern which is closest to the decision hyperplane is selected. However, in the case of remote sensing image classification with SVM, the inclusion of a single sample per iteration is not reasonable. To include several samples per iteration, we implemented the MS approach by considering the multiclass OAA SVM architecture. For each binary SVM, at each iteration, the batch of q uncertain samples closest to the decision hyperplane are selected (thus, $h \leq q \times n$ unlabeled samples are selected). The MS-cSV approach considers the smallest distance of the unlabeled samples to the n decision hyperplanes (associated to the n binary SVMs) as the uncertainty value. At each iteration, the h most uncertain samples (which do not share the closest support vector) are added to the training set. The EQB selects the h most uncertain samples according to the maximum disagreement between a committee of classifiers. The results of EQB are obtained by fixing the number of EQB predictors to eight and selecting bootstrap samples containing 75% of the initial training patterns.

The multiclass SVM with the standard OAA architecture has been implemented using the LIBSVM library (for Matlab interface) [32]. All the active-learning algorithms presented in this paper have been implemented in Matlab.

To show the effectiveness of the proposed technique, in the next section, we present the results of five different experiments. In the first and second experiments, we compared the accuracy of the proposed technique with those of the other aforementioned techniques by using one toy data set and two real data sets. The third experiment shows the robustness of the proposed approach when biased initial training samples are considered. The computational load of the different methods is analyzed in the fourth experiment. Finally, the fifth experiment shows the accuracy of the proposed technique by varying the batch size.

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

Itr No	Training Samples	$\overline{OA}$				
		Proposed	**RS**	**MS**	**MS-cSV**	**EQB**
0	3	97.43	97.43	97.43	97.43	97.43
1	6	100	92.30	94.87	94.87	89.74
2	9	100	94.87	100	100	100
3	12	100	92.30	100	100	100
4	15	100	97.43	100	100	100

B. Analysis of Results

In order to understand the potential of the proposed technique, in the first experiment, we compared the different active-learning methods by using the toy data set described in the previous section. Initially, only three samples, one from each class, are chosen for the training (see Fig. 2), and three additional samples are selected at each iteration of active learning. The process is iterated four times to have 15 samples in the training set at the convergence. Fig. 3 shows the unlabeled samples (represented with circles) which are selected by different active-learning methods after the end of the first and fourth iterations. From this figure, one can see that, for example, at the initial stage of the training, the proposed technique selects samples that are more representative of the general problem than the other techniques. For a quantitative analysis, Table III reports the classification accuracy obtained by the proposed, RS, MS, MS-cSV, and EQB methods at different iterations. From that table, one can see that the proposed technique obtained 100% classification accuracy after the first iterations (i.e., by using only six labeled samples), while the other most effective techniques (i.e., the MS, the MS-cSV, and the EQB) need at least

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

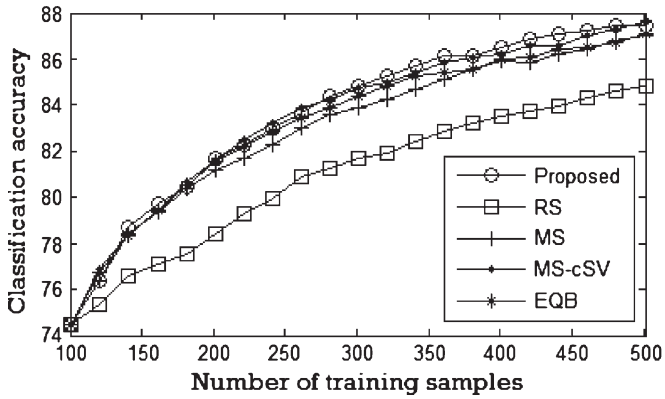

Fig. 4. Average classification accuracies over 20 runs provided by the proposed, RS, MS, MS-cSV, and EQB methods for the Paneveggio data set.

TABLE IV
AVERAGE OVERALL CLASSIFICATION ACCURACY ($\overline{OA}$), ITS STANDARD DEVIATION (s), AND KAPPA ACCURACY OBTAINED ON 20 RUNS FOR DIFFERENT TRAINING DATA SIZES (PANEVEGGIO DATA SET)

Methods	$\lvert L\rvert = 361$			$\lvert L\rvert = 421$			$\lvert L\rvert = 501$		
	$\overline{OA}$	s	kappa	$\overline{OA}$	s	kappa	$\overline{OA}$	s	kappa
Proposed	86.12	1.01	.792	86.87	0.91	.803	87.48	1.03	.812
RS	82.83	1.87	.742	83.74	1.90	.756	84.85	1.47	.772
MS	85.10	1.55	.778	85.87	1.92	.789	86.99	1.43	.806
MS-cSV	85.84	1.41	.788	86.56	1.75	.799	87.47	1.45	.812
EQB	85.39	1.57	.781	86.08	1.39	.792	87.07	1.33	.807

two iterations (i.e., nine samples) to achieve the same accuracy. In other words, although this is a simple example, starting from a suboptimal training set, the proposed technique, owing to the low-density criterion, reaches the convergence decreasing of 33% the number of new labeled samples with respect to the other literature methods.

The second experiment was carried out to compare the performance of the proposed method with those of the four techniques described in the previous section on real remote sensing data. For the Paneveggio data set, initially, only 101 labeled samples were included in the training set, and 20 samples were selected at each iteration of active learning. The whole process was iterated 20 times, resulting in 501 samples in the training set at convergence. The process was repeated for 20 trials to reduce the random effect on the results. Fig. 4 shows the average overall classification accuracies provided by different methods versus the number of samples included in the training set at different iterations. One can see that the proposed active-learning technique always produces slightly better classification accuracy than the MS method and similar accuracy compared with the MS-cSV technique. It is worth noting that, since the proposed technique selects the informative samples from the low-density regions of the kernel space, it converges faster than the MS-based approach, particularly when biased (poor) training samples are available. For a quantitative analysis, Table IV reports the mean ($\overline{OA}$), standard deviation (s), and kappa accuracies obtained on 20 runs at three different iterations (i.e., with a different number of training samples). From this table, one can seen that the standard deviation of the proposed approach is always smaller than those of the other techniques. This confirms the better stability of the proposed method versus the choice of initial training samples.

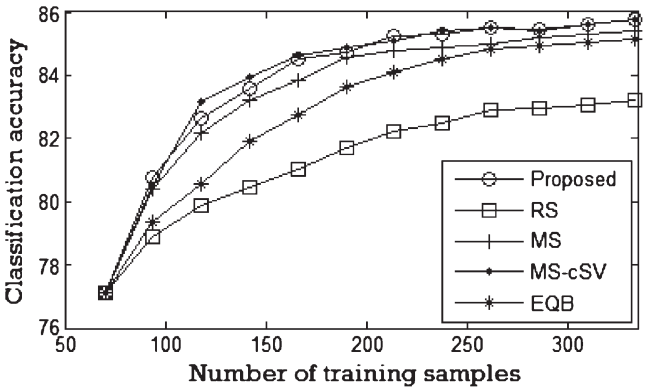

Fig. 5. Average classification accuracies over 20 runs provided by the proposed, RS, MS, MS-cSV, and EQB methods for the Pavia data set.

TABLE V
AVERAGE OVERALL CLASSIFICATION ACCURACY ($\overline{OA}$), ITS STANDARD DEVIATION (s), AND KAPPA ACCURACY OBTAINED ON 20 RUNS FOR DIFFERENT TRAINING DATA SIZES (PAVIA DATA SET)

Methods	$\lvert L\rvert = 166$			$\lvert L\rvert = 286$			$\lvert L\rvert = 334$		
	$\overline{OA}$	s	kappa	$\overline{OA}$	s	kappa	$\overline{OA}$	s	kappa
Proposed	84.50	1.00	.807	85.46	0.82	.819	85.75	0.62	.823
RS	81.00	2.29	.764	82.95	1.09	.788	83.23	1.02	.791
MS	83.86	1.74	.800	85.17	0.93	.816	85.43	0.70	.819
MS-cSV	84.65	1.12	.809	85.40	0.84	.818	85.76	0.77	.823
EQB	82.75	2.36	.786	84.94	1.26	.812	85.14	1.32	.815

For the Pavia data set, the initial training set had only 70 samples. In each round of query $h = 24$, additional samples were selected from the unlabeled pool U and added to the training set. This process of selection and training was repeated up to 11 iterations, thus including about 334 samples in the final training set. The process was repeated 20 times with different initial training samples to reduce the random effect on the results. Fig. 5 shows the average overall classification accuracy over 20 runs versus the number of samples included in the training set at different iterations. From this figure, it is clear that the proposed technique provided similar classification accuracy compared with the best technique (MS-cSV) and also converged faster than the MS and EQB methods. For a quantitative analysis, Table V reports the average overall accuracy, its standard deviation, and the kappa accuracies obtained over 20 runs at different iterations of the aforementioned learning process. From this table, one can observe that the standard deviation of the proposed approach is always smaller or comparable to the best one. This confirms also on this data set the stability of the proposed method with respect to the choice of the initial training samples.

As mentioned earlier, most of the active-learning approaches select the uncertain samples depending on the current decision hyperplane. If the initial training samples are biased, i.e., they do not provide precise representation of the classification problem, then they may fail to select proper informative samples at the initial stage of learning. On the other hand, the proposed technique selects the uncertain samples from the low-density region in the kernel space (according to the cluster assumption), and thus, it is less dependent on the quality of the initial training samples. To show the validity of the aforementioned statement, in the third experiment, we started the active-learning process with biased initial training samples. To this end, for

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

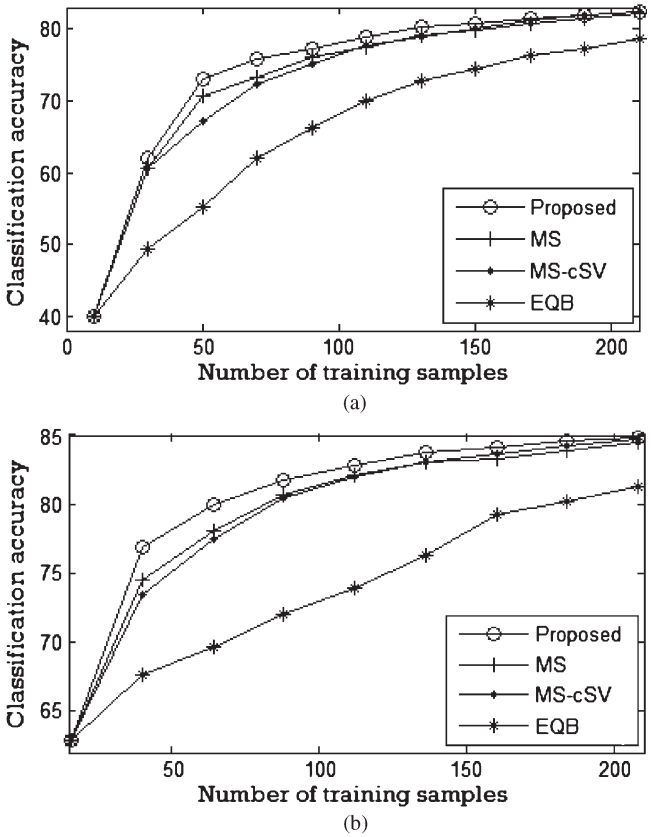

Fig. 6. Average classification accuracies provided by the proposed, RS, MS, MS-cSV, and EQB methods for the (a) Paneveggio and (b) Pavia data sets by starting with biased labeled samples.

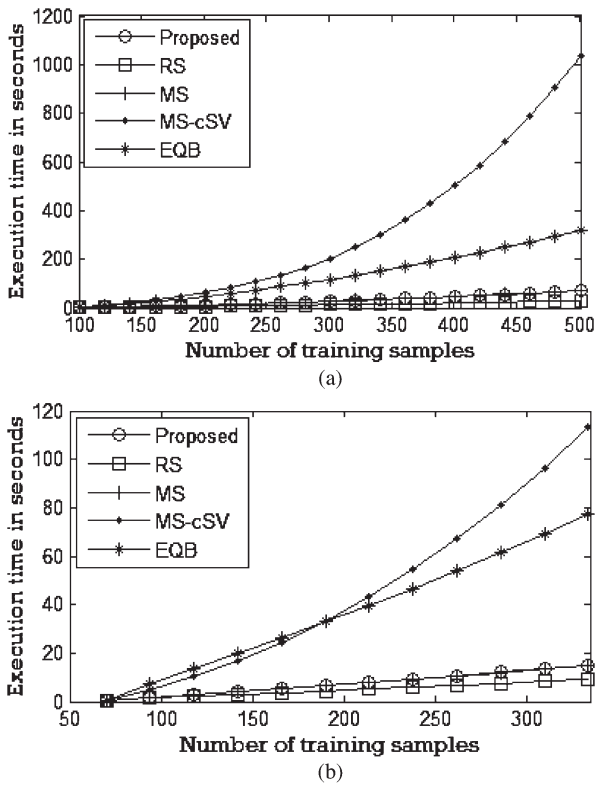

Fig. 7. Computational times taken by the proposed, RS, MS, MS-cSV, and EQB techniques at each iteration for the (a) Paneveggio and (b) Pavia data sets.

the Paneveggio and Pavia data sets, the initial training sets were defined by taking only 10 and 16 labeled samples (two samples for each class), respectively, which are not sufficient to model the actual decision boundary of the classifier. Please note that, in this experiment, our aim is to show that, during the first few iterations, the proposed technique based on the cluster assumption performs much better than the other techniques when the available initial training samples are biased, and not to solve the classification problem by considering such a few samples. Thus, for simplicity, the SVM parameters computed in the second experiment were used here. Fig. 6(a) and (b) shows the average classification accuracies versus the number of samples included in the training set at each iteration obtained by different methods for the Paneveggio and Pavia data sets, respectively. From these figures, one can see that the proposed technique always provided higher classification accuracies compared to the other methods at the initial stage of the learning process. From a different perspective, it can achieve the same accuracy of the other techniques with a significantly smaller number of samples. This confirms the robustness of the proposed technique to biased (poor) initial training samples.

The fourth experiment deals with the computational time required by the different techniques using the experimental setting (i.e., number of initial training samples, batch size, iteration number, etc.) described in the second experiment. All the experiments were carried out on a PC [INTEL(R) Core(TM)2 Duo 2.0 GHz with 2.0 GB of RAM]. Fig. 7(a) and (b) shows the computational time (in seconds) versus the number of training samples (i.e., of iterations) required by the proposed, RS, MS, MS-cSV, and EQB techniques for the Paneveggio and Pavia data sets, respectively. From these figures, one can see that, in our implementations (which could be further optimized but without changing the relative results), the computational time required by the proposed approach is almost similar to the computational time taken by the MS approach. On the contrary, the computational time taken by the MS-cSV and EQB techniques is higher compared to that of the proposed technique. From Fig. 7(a) and (b), one can see that, when the number of training samples increases (i.e., the number of SVs also increases), the MS-cSV technique takes much time to find out the uncertain samples which are closest to the distinct SVs. The RS method was obviously the most efficient in terms of computational load. Nonetheless, it resulted in the lowest classification accuracy.

The last experiment was devoted to analyze the performance of proposed technique by varying the value of the batch size h. To this end, for the Paneveggio data set, h was varied in the range 10, 15, 20, and 25, while for the Pavia data set, the value of h was varied in the range 16, 24, 32, and 40 (i.e., for each binary SVM, the number of selected uncertain samples q was varied in the range 2, 3, 4, and 5). Fig. 8(a) and (b) shows the classification accuracies versus the values of h obtained for the Paneveggio and Pavia data sets, respectively. From the analysis of the figures, one can conclude that the final accuracy of the proposed method does not significantly depend on the batch size. Finally, we carried out different trials for assessing the stability of the proposed technique, varying the width of the histogram bins. The results of all these trials (which are not

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

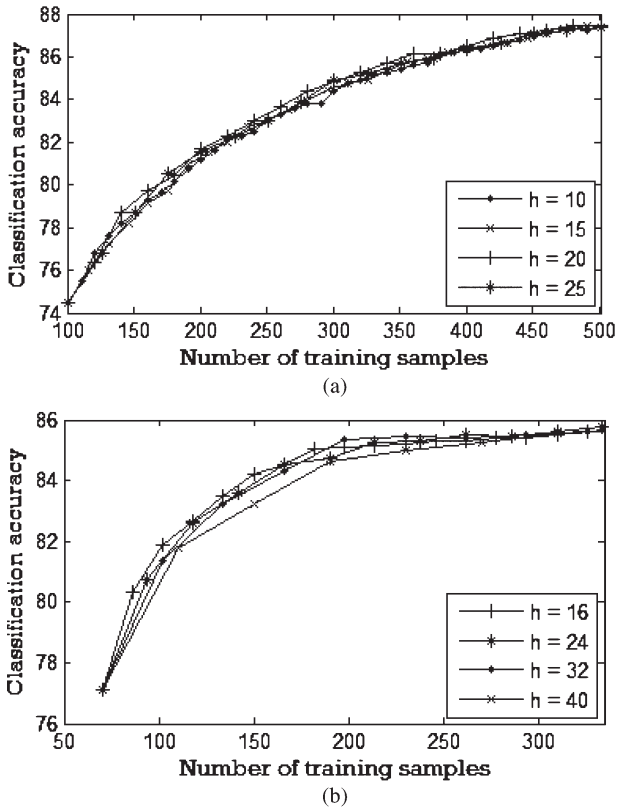

Fig. 8. Average classification accuracy provided by the proposed approach considering different values of batch size h for the (a) Paneveggio and (b) Pavia data sets.

repeated for space constraints) pointed out the insensitivity of the proposed algorithm to the width of the histogram bins.

On the basis of all the aforementioned experiments, we can conclude that, on the two considered remote sensing data sets, the proposed technique provided the best tradeoff among robustness to biased (poor) initial training samples, computational complexity, classification accuracy, and the number of new labeled samples necessary to reach convergence.

VI. Discussion and Conclusion

In this paper, we have presented a simple, fast, and reliable active-learning technique based on the cluster assumption for solving remote sensing image classification problems with SVM classifier. The proposed technique works in the 1-D output space of the SVM classifier to identify the uncertain samples. Since the classifier ranks samples from the most likely members to the most unlikely members of a class, according to the cluster assumption (which implies that the decision boundary has to lie in the low-density regions of the kernel space), the samples whose output scores fall in the valley region of the histogram are the most uncertain. Thus, the uncertain samples can be identified by finding a threshold on the histogram that identifies this valley region. Then, a batch of samples whose output scores are closest to that threshold are selected from the unlabeled pool. This makes the proposed technique relatively less dependent on both the choice of the initial training samples and the classification results at the previous iteration. This also involves faster convergence than the other techniques. It is worth noting that the robustness to biased (poor) training samples is a significant advantage in remote sensing problem where often available initial training samples do not model precisely the classification problem.

In the proposed technique we transform the original feature space into a 1-D space, thus simplifying the query function computation, which is based on looking for a threshold in the SVM output space. Thus, compared with existing methods, the proposed method is also efficient in terms of computational complexity. In addition, it can be applied to both binary and multiclass problems.

To empirically assess the effectiveness of the proposed method, we compared it with other active-learning approaches existing in the remote sensing literature using a toy data set and both a hyperspectral image and a multispectral image. By this comparison, we observed that the proposed method provides comparable accuracy to those achieved by the most effective techniques presented in the remote sensing literature (i.e., the MS-cSV and EQB methods) but with an increased robustness to biased initial training samples and a sharp reduction of the computational time. Thus, in our experiments, the proposed algorithm provided the best tradeoff among robustness to biased initial training samples, computational complexity, classification accuracy, and the number of new labeled samples necessary to reach convergence.

As a final remark, we point out that, although the performances of the proposed method were satisfactory, the method does not include any diversity criterion for selecting multiple samples. Thus, an interesting future activity would be to design the proposed query function by considering also a diversity criterion [26]. This should be done by defining a diversity criterion that can be implemented in a fast algorithm for avoiding to lose one of the most important properties of the proposed method, which is the low computational load.

Acknowledgment

The authors would like to thank the anonymous referees for their constructive criticism and valuable suggestions, and M. Dalponte and M. Tononi for the preprocessing (ground survey and data set definition) of the hyperspectral image used in this paper. This work was carried out in the framework of the India–Trento Program for Advanced Research.

References

[1] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. 16th ICML*, 1999, pp. 200–209.

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

[3] D. J. C. MacKay, "Information-based objective functions for active data selection," *Neural Comput.*, vol. 4, no. 4, pp. 590–604, Jul. 1992.

[4] P. Rigollet, "Generalization error bounds in semi-supervised classification under the cluster assumption," *J. Mach. Learn. Res.*, vol. 8, pp. 1369–1392, 2007.

[5] J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Comput. Vis. Graph. Image Process.*, vol. 29, no. 3, pp. 273–285, Mar. 1985.

[6] M. Li and I. K. Sethi, "Confidence-based active learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1251–1261, Aug. 2006.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[7] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proc. 17th Annu. Int. ACM-SIGIR Conf. Res. Develop. Inf. Retrieval*, 1994, pp. 3–12.
[8] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *J. Artif. Intell. Res.*, vol. 4, no. 1, pp. 129–145, 1996.
[9] K. Fukumizu, "Statistical active learning in multilayer perceptrons," *IEEE Trans. Neural Netw.*, vol. 11, no. 1, pp. 17–26, Jan. 2000.
[10] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proc. 5th Annu. ACM Workshop Comput. Learn. Theory*, 1992, pp. 287–294.
[11] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Mach. Learn.*, vol. 28, no. 2/3, pp. 133–168, 1997.
[12] I. Dagan and S. P. Engelson, "Committee-based sampling for training probabilistic classifiers," in *Proc. 12th ICML*, 1995, pp. 150–157.
[13] N. Abe and H. Mamitsuka, "Query learning strategies using boosting and bagging," in *Proc. 15th ICML*, 1998, pp. 1–9.
[14] P. Melville and R. J. Mooney, "Diverse ensembles for active learning," in *Proc. 21st ICML*, 2004, pp. 584–591.
[15] C. Campbell, N. Cristianini, and A. Smola, "Query learning with large margin classifiers," in *Proc. 17th ICML*, 2000, pp. 111–118.
[16] G. Schohn and D. A. Cohn, "Less is more: Active learning with support vector machines," in *Proc. 17th ICML*, 2000, pp. 839–846.
[17] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, no. 1, pp. 45–66, 2002.
[18] T. Luo, K. Kramer, D. B. Goldgof, L. O. Hall, S. Samson, A. Remsen, and T. Hopkins, "Active learning to recognize multiple types of plankton," *J. Mach. Learn. Res.*, vol. 6, no. 1, pp. 589–613, 2005.
[19] P. Mitra, B. U. Shankar, and S. K. Pal, "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern Recognit. Lett.*, vol. 25, no. 9, pp. 1067–1074, Jul. 2004.
[20] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang, "Representative sampling for text classification using support vector machines," in *Proc. 25th Eur. Conf. Inf. Retrieval Res.*, 2003, pp. 393–407.
[21] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop Comput. Learn. Theory*, 1992, pp. 144–152.
[22] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York: Springer-Verlag, 2001.
[23] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.
[24] P. Mitra, C. A. Murthy, and S. K. Pal, "A probabilistic active support vector learning algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 3, pp. 413–418, Mar. 2004.
[25] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," in *Proc. 18th ICML*, 2001, pp. 441–448.
[26] K. Brinker, "Incorporating diversity in active learning with support vector machines," in *Proc. 20th ICML*, 2003, pp. 59–66.
[27] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *Proc. 21st ICML*, 2004, pp. 623–630.
[28] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2218–2232, Jul. 2009.
[29] S. Rajan, J. Ghosh, and M. M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 4, pp. 1231–1242, Apr. 2008.
[30] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.
[31] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *J. Electron. Imaging*, vol. 13, no. 1, pp. 146–165, 2004.
[32] C.-C. Chang and C.-J. Lin, LIBSVM: A Library for Support Vector Machine, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

Swarnajyoti Patra received the B.Sc. degree in computer science and the M.C.A. degree from Vidyasagar University, Midnapur, India, in 1999 and 2003, respectively, and the Ph.D. degree in engineering from Jadavpur University, Kolkata, India, in 2009.

He is currently a Postdoctoral Researcher under the India–Trento Program for Advanced Research and Telecommunications Project with the Department of Information Engineering and Computer Science, University of Trento, Trento, Italy. His research interests include pattern recognition, evolutionary computation, neural networks, and remote sensing image analysis.

Lorenzo Bruzzone (S'95–M'98–SM'03–F'10) received the Laurea (M.S.) (*summa cum laude*) degree in electronic engineering and the Ph.D. degree in telecommunications from the University of Genoa, Genova, Italy, in 1993 and 1998, respectively.

He is currently a full Professor of telecommunications with the University of Trento, Italy, where he is the Head of the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, and teaches remote sensing, pattern recognition, radar, and electrical communications. He is the author (or coauthor) of 95 scientific publications in refereed international journals (63 in IEEE journals), more than 140 papers in conference proceedings, and 13 book chapters. He is the Editor/Coeditor of ten books/conference proceedings and one scientific book. His current research interests are in the areas of remote sensing, radar and synthetic aperture radar, signal processing, and pattern recognition (analysis of multitemporal images, feature extraction and selection, classification, object detection, regression and estimation, data fusion, and machine learning). He conducts and supervises research on these topics within the frameworks of several national and international projects.

Dr. Bruzzone is a member of the International Association for Pattern Recognition and the Italian Association for Remote Sensing (AIT). He was the General Chair and the Cochair of the First and Second IEEE International Workshop on the Analysis of Multitemporal Remote-Sensing Images and is currently a member of the Permanent Steering Committee of this series of workshops. Since 2003, he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. From 2004 to 2006, he served as an Associate Editor of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He is currently an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) and was a Guest Coeditor of different Special Issues of the IEEE TGRS. He has been a Referee of many international journals and has served on the Scientific Committees of several international conferences. He is a member of the Managing Committee of the Italian Inter-University Consortium on Telecommunications and the Scientific Committee of the India–Italy Center for Advanced Research. In 2008, he was appointed as a member of the joint NASA/ESA Science Definition Team for the radar instruments for *Outer Planet Flagship Missions*. Since 2009, he has been a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society. Since April 2010, he has been the Editor of the IEEE GEOSCIENCE AND REMOTE SENSING NEWSLETTER. He ranked first place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (Seattle, July 1998). He was the recipient of the Recognition of TGRS Best Reviewers in 1999. In the past years, joint papers presented by his students at international symposia and master theses that he supervised have received international and national awards.