Active Learning Methods for Remote Sensing Image Classification

Devis Tuia, *Student Member, IEEE*, Frédéric Ratle, Fabio Pacifici, *Student Member, IEEE*, Mikhail F. Kanevski, and William J. Emery, *Fellow, IEEE*

***Abstract*—In this paper, we propose two active learning algorithms for semiautomatic definition of training samples in remote sensing image classification. Based on predefined heuristics, the classifier ranks the unlabeled pixels and automatically chooses those that are considered the most valuable for its improvement. Once the pixels have been selected, the analyst labels them manually and the process is iterated. Starting with a small and nonoptimal training set, the model itself builds the optimal set of samples which minimizes the classification error. We have applied the proposed algorithms to a variety of remote sensing data, including very high resolution and hyperspectral images, using support vector machines. Experimental results confirm the consistency of the methods. The required number of training samples can be reduced to 10% using the methods proposed, reaching the same level of accuracy as larger data sets. A comparison with a state-of-the-art active learning method, margin sampling, is provided, highlighting advantages of the methods proposed. The effect of spatial resolution and separability of the classes on the quality of the selection of pixels is also discussed.**

***Index Terms*—Active learning, entropy, hyperspectral imagery, image information mining, margin sampling (MS), query learning, support vector machines (SVMs), very high resolution (VHR) imagery.**

I. Introduction

With the increase of spatial and spectral resolution of recently launched satellites, new opportunities to use remotely sensed data have arisen. Fine spatial optical sensors with metric or submetric resolution, such as QuickBird, Ikonos, WorldView-1, or the future WorldView-2 mission, allow detecting fine-scale objects such as elements of residential housing, commercial buildings, and transportation systems and utilities. Hyperspectral remote sensing systems provide additional discriminative features for classes that are spectrally similar, due to their higher spectral resolution.

All these information sources will generate large archives of images, creating a need for automatic procedures for information mining. Kernel methods and, in particular, support vector machines (SVMs) [1], [2] have shown excellent performance in multispectral [3]–[5] and hyperspectral [6]–[10] image classification, owing to the following: 1) their valuable generalization properties; 2) their ability to handle high-dimensional feature spaces; and 3) the uniqueness of the solution.

Despite their excellent performances, SVMs, as any supervised classifier, rely on the quality of the labeled data used for training. Therefore, the training samples should be fully representative of the surface-type statistics to allow the classifier to find the correct solution. This constraint makes the generation of an appropriate training set a difficult and expensive task which requires extensive manual (and often subjective) human-image interaction. Manual training set definition is usually done by visual inspection of the scene and the successive labeling of each sample. This phase is highly redundant as well as time-consuming. In fact, several neighboring pixels carrying the same information are included in the training set. Such a redundancy, although not harmful for the quality of the results if performed correctly, slows down the training phase considerably. Therefore, in order to make the models as efficient as possible, the training set should be kept as small as possible and focused on the pixels that really help to improve the performance of the model. This is particularly important for very high resolution (VHR) images that may easily reach several millions of pixels. There is a need for procedures that automatically, or semiautomatically, define a suitable (in terms of information and computational costs) training set for satellite image classification, particularly in the context of complex scenes such as urban areas.

In the machine learning literature, this problem is known as *active learning*. A predictor trained on a small set of well-chosen examples can perform as efficiently as a predictor trained on a larger number of examples randomly chosen, while being computationally smaller [12]–[14]. Following this idea, active learning exploits the user–machine interaction, decreasing simultaneously both the classifier error, using an optimized training set, and the user's effort to build this set. Such a procedure, starting with a small and nonoptimal training set, presents to the user the pixels whose inclusion in the set improves the performance of the classifier. The user interacts with the machine by labeling such pixels. The procedure is then iterated until a stopping criterion is met.

In the active learning paradigm, the model has control over the selection of training examples between a list of candidates. This control is given by a problem-dependent heuristic, for

Manuscript received August 28, 2008; revised October 27, 2008 and November 12, 2008. First published April 7, 2009; current version published June 19, 2009. This work was supported in part by the Swiss National Foundation under Grants 100012-113506, 105211-107862, and 200021-113944.

D. Tuia, F. Ratle, and M. F. Kanevski are with the Institute of Geomatics and Analysis of Risk, University of Lausanne, 1015 Lausanne, Switzerland.

F. Pacifici is with the Earth Observation Laboratory, Department of Computer, Systems and Production Engineering, "Tor Vergata" University, 00133 Rome, Italy.

W. J. Emery is with the Department of Aerospace Engineering Sciences, University of Colorado, Boulder, CO 80309 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2008.2010404

instance, the decrease in test error if a given candidate is added to the training set [13]. The active learning framework is effective when applied to learning problems with large amounts of data. This is the case of remote sensing images, for which active learning methods are particularly relevant since, as stated earlier, the manual definition of a suitable training set is generally costly and redundant. Several active learning methods have been proposed so far. They may be grouped into three different classes, which will be briefly discussed next.

The first class of active learning methods relies on SVM specificities [15]–[17] and has been widely applied in environmental science for monitoring network optimization [18], species recognition [19], in computer vision for image retrieval [20]–[24], and in linguistics for text classification and retrieval [25], [26]. These active methods take advantage of the geometrical features of SVMs. For instance, the margin sampling (MS) strategy [15], [16] samples the candidates lying within the margin of the current SVM by computing their distance to the dividing hyperplane. This way, the probability of sampling a candidate that will become a support vector is maximized. Tong and Koller [25] proved the efficiency of these methods. Mitra *et al.* [27] discussed the robustness of the method and proposed confidence factors to measure the closeness of the SVM found to the optimal SVM. Recently, Ferecatu and Boujemaa [24] proposed to add a constraint of orthogonality to the MS, resulting in maximal distance between the chosen examples.

The second class of active learning methods relies on the estimation of the posterior probability distribution function (pdf) of the classes, i.e., $p(\cdot|\cdot)$. The posterior distribution is estimated for the current classifier and then confronted with n data distributions, one for each of the n candidate points individually added to the current training set. Thus, as many posterior pdfs as there are unknown examples have to be estimated. In [28], uncertainty sampling is computed for a two-class problem: the selected samples are the ones giving the class membership probability closest to 0.5. In [29], a multiclass algorithm is proposed. The candidate that maximizes the Kullback–Leibler divergence (or relative entropy KL [30]) between the distributions is added to the training set. These methods can be adapted to any classifier giving probabilistic outputs but are not well suited for SVM classification, given the high computational cost involved.

The last class of active methods is based on the query-by-committee paradigm [31]–[33]. A committee of classifiers using different hypotheses about parameters is trained to label a set of unknown examples (the candidates). The algorithm selects the samples where the disagreement between the classifiers is maximal. The number of hypotheses to cover becomes quickly computationally intractable for real applications [34], and approaches based on multiple classifier systems have been proposed [35]. In [36], methods based on boosting [37] and bagging [38] are described as adaptations of the query-by-committee. In [36], the problem is applied solely to binary classification. In [39], results obtained by query-by-boosting and query-by-bagging are compared on several batch data sets showing excellent performance of the methods proposed. In [40], expectation–maximization and a probabilistic active learning method based on query-by-committee are combined for text classification: in this application, the disagreement between classifiers is computed by the KL divergence between the posterior pdf of each member of the committee and the mean posterior distribution function.

Despite both their theoretical and experimental advantages, active learning methods can rarely be found in remote sensing image classification. Mitra *et al.* [41] discussed an SVM MS method similar to that in [15] for object-oriented classification. The method was applied successfully to a 512 × 512 multispectral four-band image of the Indian Remote Sensing satellite with a spatial resolution of 36.25 m. Only a single pixel was added at each iteration, requiring several retrainings of the SVM, resulting in high computational cost. Rajan *et al.* [42] proposed a probabilistic method based on the study in [29] using maximum likelihood classifiers for pixel-based classification. This method showed excellent performances on two data sets. The first was a 512 × 614 NASA Airborne Visible/Infrared Imaging spectrometer at 18-m resolution, while the second was a Hyperion 1476 × 256 image (30 m of spatial resolution). Unfortunately, the approach proposed cannot be applied to SVMs, again, because of their computational cost. Recently, Jun and Ghosh [43] extended this approach, proposing to use boosting to weight pixels that were previously selected but were no longer relevant for the current classifier. Zhang *et al.* [44] proposed information-based active learning for target detection of buried objects. More recently, this approach was extended by Liu *et al.* [45], who proposed a semisupervised method based on active queries: in this study, the advantages of active learning to label pixels and of semisupervised learning to exploit the structure of unlabeled data are fused to improve the detection of targets.

In this paper, we propose two variations of active learning models that aim at improving the adaptability and speed of the existing methods. The first algorithm, the MS by closest support vector (MS-cSV, Section II-B), is an extension of MS [15] and aims at solving the problem of simultaneous selection of several candidates addressed in [24]: the original heuristic of MS is optimal when a single candidate is chosen at every iteration. When several samples are chosen simultaneously, their distribution in the feature space is not considered, and therefore, several samples lying in the same region close to the hyperplane, i.e., possibly providing the same information, are added to the training set. We propose a modification of the MS heuristic in order to take this effect into account. In fact, by adding a constraint on the distribution of the candidates, only one candidate per region of the feature space is sampled. Such a modification allows sampling of several candidates at every iteration, improving the speed of the algorithm and conserving its performance. The second algorithm, the entropy query-by-bagging (EQB, Section II-C), is an extension of the query-by-bagging algorithm presented in [36]. In order to obtain a multiclass extension of this algorithm, we exploit an entropy-based heuristic. The disagreement between members of the committee of learners is therefore expressed in terms of entropy in the distribution of the labels provided by the members. A candidate showing maximum entropy between the predictions is poorly handled by the current classifier and is therefore added to the training set. Since this approach belongs to the query-by-committee algorithms, it has the fundamental advantage of being independent from the classifier used and can be applied when using other methods, such as neural networks. In this paper, SVM classifiers have been used to provide a fair comparison with the other methods.

Both methods are compared to classical MS (briefly recalled in Section II-A) on three different test cases, including VHR optical imagery and hyperspectral data. The VHR optical imagery consists of two QuickBird data sets. The first image is a scene of Rome (Italy) imaged at 2.4-m resolution, while the second pansharpened mulstispectral image was acquired over Las Vegas (Nevada, U.S.) at 0.6-m resolution. The hyperspectral data are an AVIRIS image of the Kennedy Space Center (KSC, Florida, U.S.) at the spatial resolution of 18 m, made available for comparison from [46]. For each data set, the algorithm starts with a small number of labeled pixels and adds pixels iteratively from the list of candidates. In order to avoid manual labeling at each iteration, the candidates have been previously labeled (ground survey). Error bounds are provided by an algorithm adding the candidates randomly (upper) and by an SVM trained on the complete ground survey (lower).

This paper is organized as follows: Section II presents the MS approach and the two algorithms proposed. Section III presents the data sets, while the experimental results are discussed in Section IV. Final conclusions are in Section V.

II. Active Learning Algorithms

Consider a synthetic illustrative example (Fig. 1). We have a training set composed by n labeled examples consisting of a set of points $X = \{x_1, x_2, \ldots, x_n\}$ and corresponding labels $Y = \{y_1, y_2, \ldots, y_n\}$ [Fig. 1(a)]. We wish to add to the training set a series of examples from a set of m unlabeled points $Q = \{q_1, q_2, \ldots, q_m\}$ [Fig. 1(b)], with $m \gg n$. X and Q have the same features. The examples are not chosen randomly, but by following a problem-oriented heuristic that aims at maximizing the performances of the classifiers. Fig. 1(c) shows the training set obtained by random selection of points. Fig. 1(d) shows the training set obtained using an active learning method. In this case, the algorithm concentrates on difficult examples, i.e., the examples lying on the boundaries between classes. This is due to the fact that the classifier has control over the sampling and avoids taking examples in regions that are already well classified; the classifier favors examples that lie in regions of high uncertainty.

These considerations hold when $p(y|x)$ is smooth and the noise can be neglected. In the case of very noisy data, an active learning algorithm might include in the training set noisy and uninformative examples, resulting in a selection equivalent to random sampling. In remote sensing, such an assumption about noise holds for multispectral and hyperspectral imagery, but it does not for synthetic aperture radar imagery, where the algorithms discussed hereafter can hardly be applied. In this paper, we will focus on data that satisfy these assumptions.

As stated in the Introduction, different strategies have been proposed in the literature for the active selection of training examples. The following sections present the MS algorithm and the active learning approaches proposed in this paper.

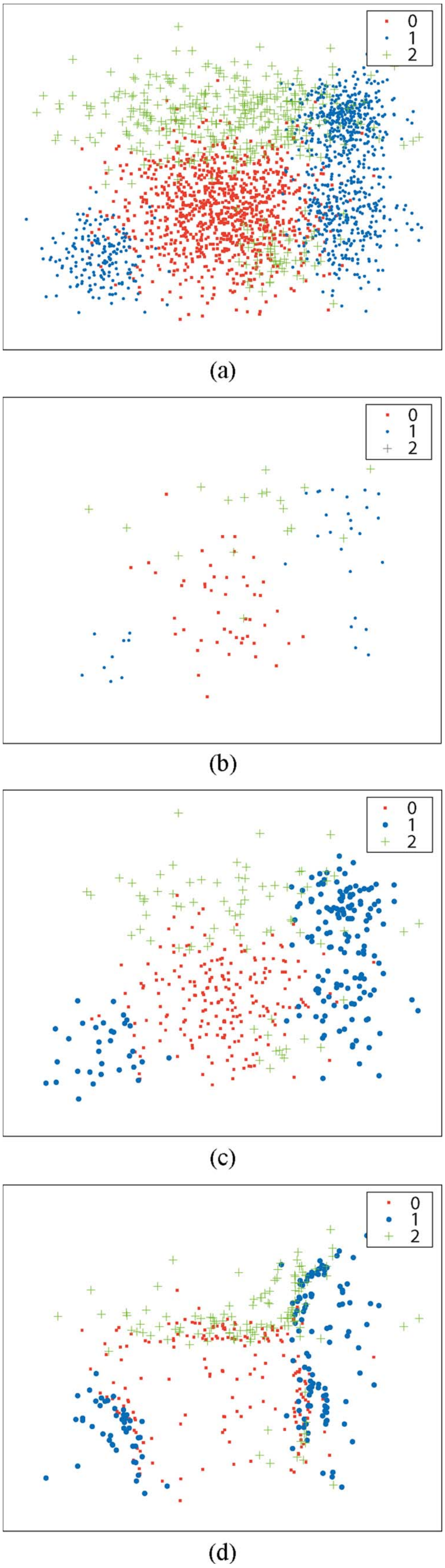


Fig. 1. Example of active selection of training pixels. (a) Initial training set X (labeled). (b) Unlabeled candidates Q. (c) Random selection of training examples. (d) Active selection of training examples: the training examples are chosen along the decision boundaries.

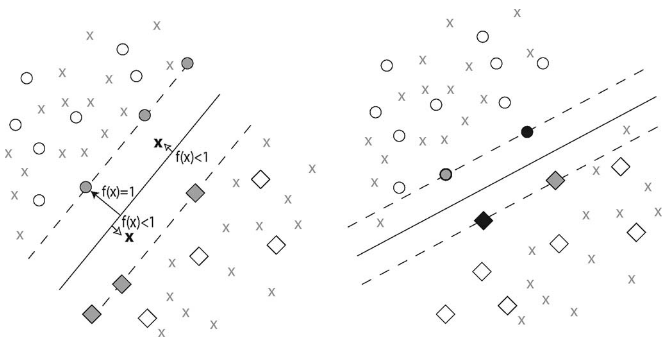


Fig. 2. MS active learning. (Left) SVM before inclusion of the two most interesting examples. (Right) New SVM decision boundary after inclusion of the new training examples.

A. MS

MS is a SVM-specific active learning algorithm taking advantage of SVM geometrical properties [15]: assuming a linearly separable case, where the two classes are separated by a hyperplane given by the SVM classifier [Fig. 2(a)], the support vectors are the labeled examples that lie on the margin at a distance of exactly 1 from the decision boundary (filled circles and diamonds in Fig. 2). If we now consider an ensemble of unlabeled candidates ("X"s in Fig. 2), we make the assumption that the most interesting candidates are the ones that fall within the margin of the current classifier, as they are the most likely to become new support vectors [Fig. 2(b)].

Consider the decision function of the two-class SVM

$$f(q_i) = \text{sign}\left(\sum_{j=1}^{n} y_j \alpha_j K(x_j, q_i)\right) \tag{1}$$

where $K(x_j, q_i)$ is the kernel matrix, which defines the similarity between the candidate q_i and the j-th support vector; α are the support vector coefficients; and y_j's are their labels of the form $\{\pm 1\}$. In a multiclass context and using a one-against-all SVM [2], a separate classifier is trained for each class cl against all the others, giving a class-specific decision function $f_{cl}(q_i)$. The class attributed to the candidate q_i is the one minimizing $f_{cl}(q_i)$.

Therefore, the candidate included in the training set is the one that respects the condition

$$\hat{x} = \arg\min_{q_i \in Q} |f(q_i)| . \tag{2}$$

In the case of remote sensing imagery classified with SVM, the inclusion of a single candidate per iteration is not optimal. Considering computational cost of the model (cubic with respect to the observations), inclusion of several candidates per iteration is preferable. MS (detailed in Algorithm 1) provides a set of candidates at every iteration. However, MS has not been designed for this purpose, and such a straightforward adaptation of the method is not optimal on its own. The left side of Fig. 3 shows the effect of a nonuniform distribution of candidates when several neighboring examples lie close to the margin: if the MS algorithm chooses three examples in a single run, three candidates from the same neighborhood will be chosen. To avoid such a problem, we propose the MS-cSV algorithm that will be addressed in the next section.

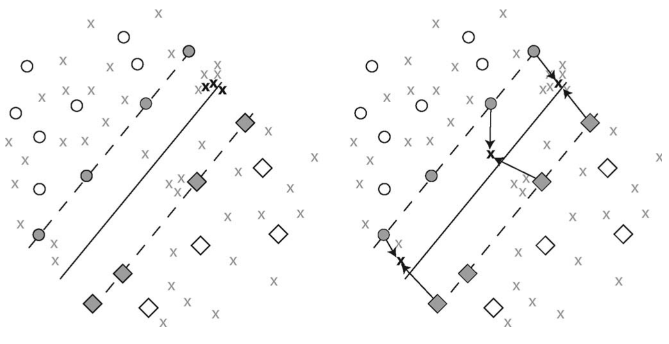


Fig. 3. MS active learning. (Left) Candidates chosen by the MS. (Right) Candidates chosen taking into account the support vector distribution.

B. MS-cSV

As stated earlier, one of the drawbacks of the MS is that the method is optimal only when a single candidate is chosen per iteration. In order to take into account the distribution in the feature space of the candidates, we propose a modification of the margin sample algorithm. The position of each candidate with respect to the current support vectors is stored, and this information is used to choose the most interesting examples.

The SVM solution provides a list of support vectors $SV = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ with $\alpha \neq 0$. For every candidate q_i, we can select the closest support vector $cSV = \arg\min_{x_j \in SV} K(x_j, q_i)$.

The heuristic of (2) can be modified in order to include an additional constraint: when confronted with several candidates located in the vicinity of the same support vector, the algorithm only takes into account the candidate associated with the minimal distance to the margin. In other words, no points added can share the closest support vector at each iteration, as shown by

$$\hat{x}_h = \arg\min_{q_i \in Q} \left(|f(q_i)|\right) \cap cSV_h \neq cSV_l \tag{3}$$

where $l = 1, \ldots, h-1$ are the indices of the already selected candidates. Algorithm 2 presents the MS-cSV.

It is important to notice that if we add only one candidate at each iteration, the MS-cSV algorithm is identical to MS. Adding only one point makes the cSV constraint useless, since the point chosen is the one minimizing the distance to the margin between all the unique regions in the feature space: this point is simply the one minimizing the distance to the margin over the candidates.

C. EQB

The query-by-bagging approach is quite different from the approaches discussed previously. As stated in the Introduction, the algorithm belongs to the query-by-committee algorithms, for which the choice of a candidate is based on the maximum disagreement between a committee of classifiers. In the implementation of the approach in [36], bagging (*bootstrap aggregation* [38]) is proposed to build the committee: first, k training sets built on bootstrap samples [48], i.e., a draw with replacement of the original data, are defined. Then, each set is used to train a SVM classifier and to predict the class membership of the m candidates. At the end of the bagging procedure, k possible labelings of each candidate are provided.

The approach proposed in [36] has been discussed for binary classification: the candidates that will be added to the training set are the ones for which the predictions are the most evenly split, as shown in

$$\hat{x} = \arg\min_{q_i \in Q} \| \{t \leq k | f_t(q_i) = 1\} | - | \{t \leq k | f_t(q_i) = 0\} \| \tag{4}$$

where t is one of the k classifiers and the binary labels are of the form $\{0, 1\}$. If the classifiers agree to a certain classification, (4) is maximized. On the contrary, uncertain candidates yield small values.

In the implementation proposed in this paper, the heuristic of (4) is replaced by a multiclass one based on the maximum entropy of the distribution of the predictions of the k classifiers [see (6)]. By considering the k labels of a given candidate q_i, it is possible to compute the entropy of the distribution of the labels $H(q_i)$ using

$$H(q_i) = \sum_{cl} -p_{i,cl} \log(p_{i,cl}) \tag{5}$$

Where $p_{i,cl}$ is the probability to have the class cl predicted for the candidate i. $H(q_i)$ is computed for each candidate in Q, and then, the candidates satisfying the heuristic

$$\hat{x} = \arg\max_{q_i \in Q} H(q_i) \tag{6}$$

Are added to the training set.

Entropy maximization gives a naturally multiclass heuristic. A candidate for which all the classifiers in the committee agree is associated with null entropy; such a candidate is already correctly labeled by the classifiers, and its inclusion does not bring additional information. On the contrary, a candidate with maximum disagreement between the classifiers results in maximum entropy, i.e., a situation where the predictions given by the k classifiers are the most evenly split. Therefore, the parallels with the original query-by-bagging formulation are strong.

The EQB does not depend on SVM characteristics but on the distribution of k class memberships resulting from the committee learning. Therefore, it depends on the outputs of the classifiers only and can be applied to any type of classifier (maximum likelihood, neural networks, Bayes classifiers, etc.).

Regarding computational cost of the method, some specific considerations can be done depending on the classifier used: when using a SVM, the cost remains competitive compared to the MS presented earlier, because the training phase scales linearly with respect to the number of models k (when all the training sets are drawn in the bootstrap samples) compared to the MS using the entire training set. For smaller draws in the bootstrap samples, the additional computational burden becomes smaller than linear. When using probabilistic classifiers and in comparison with models based on posterior pdf estimation, EQB implies k trainings for each iteration, instead of m trainings related to the estimation of the probability distribution for each set update with a candidate.

Therefore, simply using the entropy on the k predictions of the candidates is computationally less expensive than using the methods presented earlier. Algorithm 3 presents the EQB.

TABLE I
DATA SETS CONSIDERED

Location	Rome (Italy)	Las Vegas (Nevada, USA)	KSC (Florida, USA)
Dim. (pixels)	706 x 729	755 x 722	614 x 512
Satellite	QuickBird	QuickBird	NASA AVIRIS
Acq. Date	May 29, 2002	May 10, 2002	March 23, 1996
Spat. Res.(m)	2.4	0.6	18.0

III. Data Sets

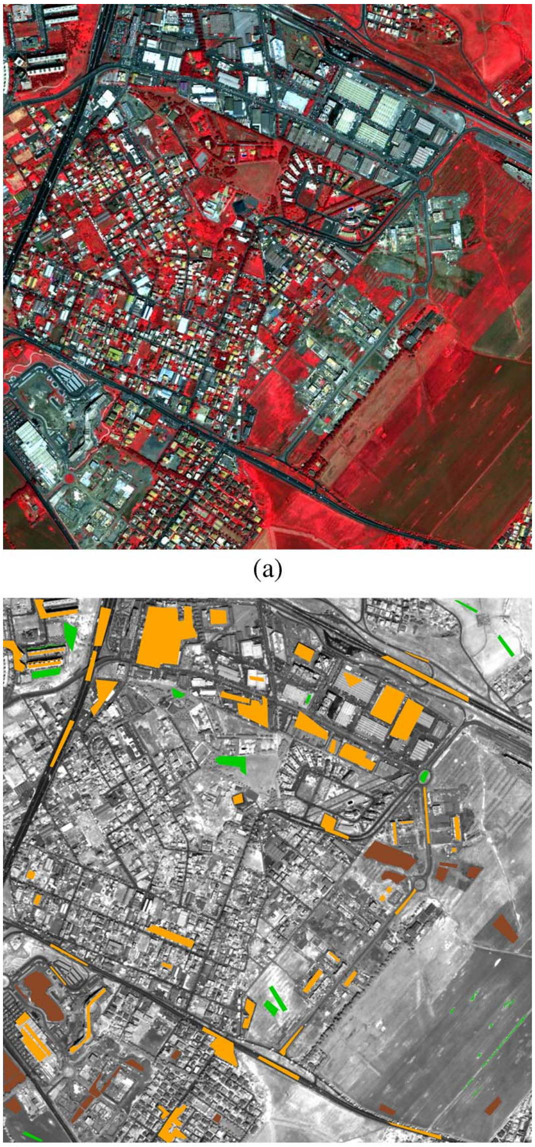
The VHR data sets used are portions of the cities of Rome (Italy) and Las Vegas (Nevada, U.S.), acquired by QuickBird in 2002 and 2004, respectively. In particular, two different spatial resolutions have been considered: 2.4 m multispectral for the Rome case and 0.6 m pansharpened multispectral for Las Vegas. Furthermore, an 18-m spatial resolution hyperspectral image of the KSC, acquired by AVIRIS in 1996, has been used for comparison and validation purposes. Details of scenes and images are reported in Table I. This variety in land covers/land uses made possible the evaluation of the flexibility of our active learning procedure when applied to different landscapes and spatial/spectral resolutions, since different surfaces of interest have been recognized with respect to the specificities of each scene.

We have chosen here to consider the pixels of the unlabeled Q set from the labeled training set ($Q = [\text{training set}] - X$) in order to avoid manual labeling between the iterations. Note that the labels of the candidates are never used in the selection process. They are attributed to the pixels when (and if) they are added to the current training set X. For future applications to unknown images, the Q set will be composed of the unlabeled examples of the image (or a part of them), and pixels selected by the machine will be labeled by the user only after their selection. The pixels of the X set will be the only labeled examples at the beginning of the active selection.

A. Rome

The Rome test site, shown in Fig. 4(a), is next to the campus of "Tor Vergata" University which is located in southeast Rome, Italy. This area is a typical suburban scene with residential, commercial, and industrial buildings. It is possible to notice the construction of several buildings, including a shopping mall in the bottom left of the image, which highlights the considerable land surface changes that this area underwent during the past decade. The different land-cover surfaces have been grouped into three main classes:

1) *man-made*, including buildings, concrete, asphalt, gravel, and sites under construction;
2) *green vegetation*;
3) *bare soil*, including low density, dry vegetation, and unproductive surfaces.



(a)

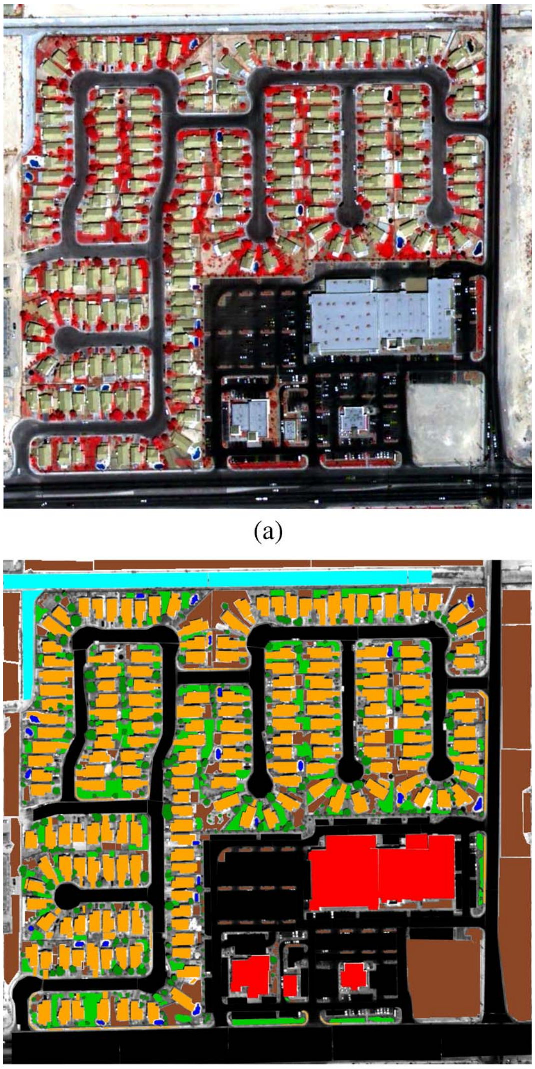
(b)

Fig. 4. (a) Rome 2.4-m QuickBird image. (b) Ground survey used (orange = man made, green = vegetation, and brown = soil).

TABLE II
CLASSES, SAMPLES OF THE GROUND SURVEY, AND LEGEND COLOR OF THE ROME DATA SET

Class	GT pixels	legend
Man made	22 318	Orange
Vegetation	2 673	Green
Soil	6 945	Brown

The reference ground survey of 31 936 pixels [Fig. 4(b)] has been randomly split in a training set (used for both X and Q sets) of 18 000 pixels, a validation set (to estimate the optimal parameters) of 7000 pixels, and a test set (to compute the test error at every step of the algorithm) of 6936 pixels. The number of labeled pixels and the reference map colors are given in Table II.



(a)

(b)

Fig. 5. (a) Las Vegas 0.6-m QuickBird image. (b) Ground survey used (orange = residential buildings, red = commercial buildings, black = asphalt, light green = short vegetation, dark green = trees, brown = soil, blue = water, and cyan = drainage channel).

The initial data set X has been set to 300 pixels, which is a small training set considering the dimensions of a VHR image. Each algorithm ran for 70 epochs, adding the 60 most relevant pixels to the actual training set at each iteration. After testing several hyperparameters for EQB sets and taking into account the computational cost, the number of predictors k has been set to eight. Every ensemble of bootstrap X'_l contains 75% of the pixels of X. In order to avoid the effects of different initializations on performance, the entire procedure has been run 16 times with different starting sets X and Q.

B. Las Vegas

The Las Vegas scene (see Fig. 5) contains regular criss-crossed roads and different examples of buildings characterized by similar heights (about one or two floors) with different dimensions, from small (residential houses) to large (commercial buildings).

TABLE III
CLASSES, SAMPLES OF THE GROUND SURVEY, AND LEGEND COLOR OF THE LAS VEGAS DATA SET

Class	GT pixels	color
Residential buildings	87590	Orange
Commercial buildings	22769	Red
Asphalt	139871	Black
Short vegetation	22414	Light green
Trees	13038	Dark Green
Soil	71582	Brown
Water	1472	Blue
Drainage channel	14287	Cyan

This second scene was chosen to represent a typical American suburban landscape, including small residential houses and large roads, different from the European style of old cities built with a more complex lattice. For instance, an unusual structure within the scene was a "Drainage Channel" located in the upper part of the image. This construction had a shape similar to roads, but brighter since it was made of concrete. A further discrimination was made between "Residential Houses" and "Commercial Buildings" due to the difference in the color of the roofs. Finally, more traditional classes such as "Trees," "Short Vegetation," and "Water" were added for a total of eight classes. Details on the number of labeled samples are reported in Table III.

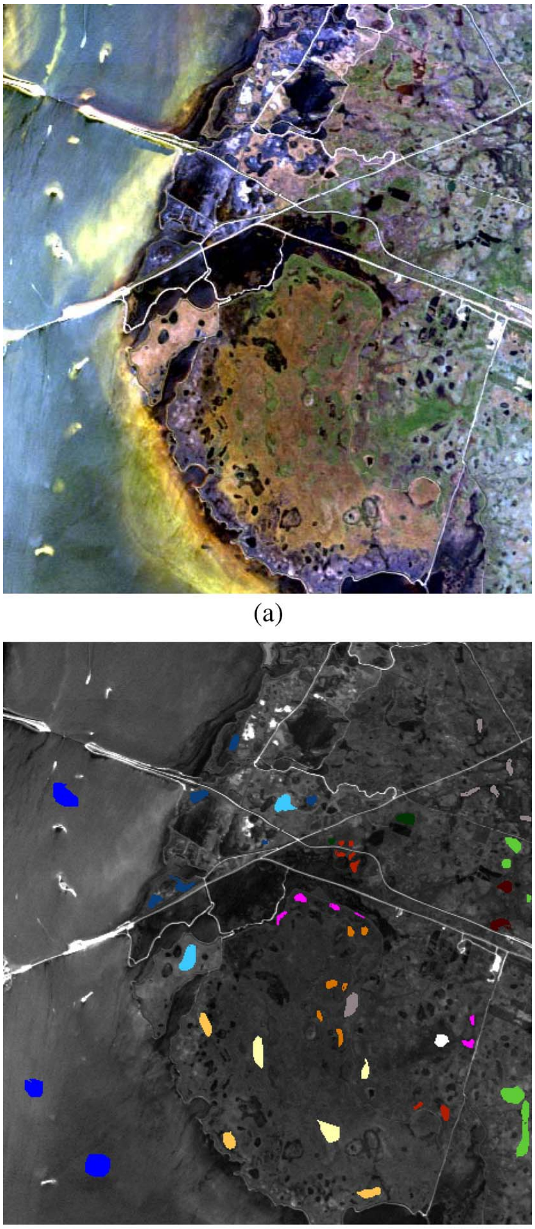
A reference ground survey of 373 023 pixels [Fig. 4(b)] has been split randomly into a training set of 30 000 pixels, a validation set of 25 000 pixels, and a test set of 318 023 pixels.

The land-use classes given in Table III have been considered. The initial data set X consists of 1000 pixels, in order to take into account enough information for all the classes. The algorithm ran for 70 epochs, adding the 80 most relevant pixels to the current training set at each iteration. Following a search of optimal EQB parameters, the number of EQB predictors k has been set to eight. Every bootstrap sample X_l' contains 75% of the pixels of X. The entire procedure has been run 11 times with different initial sets X and Q.

C. KSC, Florida

The third image (see Fig. 6), used here for comparison, has been used in [46] as a test site for hyperspectral classification. It has been acquired over the KSC, Florida, U.S., on March 23, 1996 by the hyperspectral NASA AVIRIS instrument (224 bands of 10-nm width). Water absorption and low SNR bands have been removed, resulting in a total of 176 bands. Thirteen classes representing the various land cover types were defined (see Table IV) according to the study in [46]. In that paper, the authors pointed out the similarity of spectral signatures for certain vegetation types, particularly for classes 4 and 6 that are, in fact, mixed classes.

A 5211-pixel ground survey has been split randomly into a training set of 2500 pixels, a validation set of 1300 pixels, and a test set of 1411 pixels. The initial data set X has been set to 200 pixels, which is a small training set considering the dimensions of an AVIRIS image. The algorithm ran for 70 epochs, adding the 30 most relevant pixels to the actual training set at each iteration.


(a)

(b)

Fig. 6. (a) KSC 18-m AVIRIS image. (b) Ground survey used (see Table IV for legend colors).

TABLE IV
CLASSES, SAMPLES OF THE GROUND SURVEY, AND LEGEND COLOR OF THE KSC DATA SET

Class	GT pixels	color
Scrub	761	Light green
Willow swamp	243	Pink
Cabbage palm hammock	256	Dark orange
Cabbage palm/oak hammock	252	Red
Slash pine	161	Dark green
Oak/broadleaf hammock	229	Burgandy
Hardwood swamp	105	White
Graminoid marsh	431	Gray
Spartina marsh	520	Yellow
Cattail marsh	404	Orange
Salt marsh	419	Sky blue
Mud flats	503	Steel blue
Water	927	Blue

IV. Results and Discussion

For each data set, an SVM model trained on all the ground survey pixels has been taken as a reference for the best achievable performance of the classifier ("Full SVM" thereafter). The Full SVM is taken as the lower bound of the error.

As upper bound reference, a model adding randomly $Npts$ candidates from the Q set at every iteration has been used. Since the candidates are chosen from the Q set only, the random selection performs stratified selection ("SRS" thereafter) on the labeled areas [47]. A second random sampling strategy, called spatial random sampling ("SpRS" thereafter), has been added to account for a more realistic random sampling, where the candidates are selected on a uniform spatial grid. To guarantee fair comparison, no additional stratification with respect to the distribution of the labels is done. The error related to the SRS error can be interpreted as an upper bound because all the active methods have the objective to converge to the Full SVM performance faster than the SRS of examples from the list of candidates. It is important to recall that even SRS will converge to the Full SVM error rate, but slower than the active methods.

Optimal SVM parameters $\Theta = \{\sigma, C\}$ (RBF kernel is used) are found by grid search on the parameter space. C is a parameter controlling the tradeoff between model complexity and training error. The grid search procedure allows estimating the best parameters for the initial SVM. Obviously, these parameters can become suboptimal as the training set size increases. Nonetheless, a grid search procedure implies the training and testing of as many SVMs as the sets of parameters $\Theta_{ij} = \{\sigma_i, C_j\}$ considered, and reestimating the parameters during the procedure is computationally very expensive. Thus, reestimation of parameters is when the solution seems to be trapped in a local minimum. In this paper, the reestimation has been necessary for the Las Vegas case study only, as detailed in Section IV.

The preprocessing of the images (preparation of ground survey and data extraction) has been done in ENVI 4.3. A multiclass SVM (with the one-against-all approach) has been implemented using the Torch 3 library [49]. The active learning algorithms (MS, MS-cSV, and EQB) have been implemented in Matlab 7.

A. Rome

For the Rome data set, the Full SVM achieves an overall error of 8.70% with relative kappa index of 0.810. The processing time, including model selection, is of one day on a dual-core Pentium PC (3-Gb RAM, 2.99 GHz). Fig. 7(a) shows the evolution of the test error over the iterative process for the three algorithms considered and SRS. For the Rome data set, the three active algorithms perform similarly and converge to the Full SVM error in about 25 iterations, i.e., using about 1800 training pixels. This corresponds to 10% of the training set used by the Full SVM. The MS-cSV algorithm is the one giving the best performances, owing to its higher accuracy for the class "soil." This class is the most difficult, because of its high overlap with the class "Man made" in the construction sites (see the construction site in the bottom left corner of Fig. 7). In light of the quick convergence to the Full SVM accuracy of the three active methods and keeping in mind the need for computational parsimony, no parameter reestimation is performed. Regarding computational time, the cSV model performed the first 20 iterations in 1 h (including the model selection) and ended with 35 min per iteration at iteration 70, when 4500 pixels were considered in X. Therefore, the algorithm needed approximately half an hour to converge to the optimal solution reported in Table V and remains highly competitive with respect to the Full SVM.

Accuracies per class are given in Table V for iteration 26 (1860 pixels): all the active methods outperform SRS in terms of kappa index for the three classes. The methods based on MS outperform the EQB for this example, for which EQB shows smaller accuracies than MS and MS-cSV (see the curves in Fig. 7(a) and Table V).

The classification map of the MS-cSV method [Fig. 8(b)] shows the good performance of the active algorithms (results for the MS and EQB are similar and are omitted to avoid redundancy in the figures): this map has been produced with 10% of the training examples used to produce Fig. 8(a). The soil region at the bottom right corner is where the biggest differences can be seen: this region is characterized by mixed land cover where both soil and vegetation are present. In this region, the active algorithm is still not optimal. Nonetheless, in some regions (for instance, the bottom left corner for the class "soil" and the bottom center for the class "vegetation"), the active algorithm has a tendency to suppress noise that could be generated by inconsistencies in the full training set. This is most likely related to the small size of the training set and to the active strategy: by including few pixels carefully chosen near the boundary of the classes, the redundancy in the class definition is limited and the emphasis is put on difficult pixels, i.e., pixels showing mixed spectral response.

B. Las Vegas

For the pansharpened Las Vegas image, eight classes have to be discriminated and the composition of the training set is highly unbalanced (the class "water" has only 1472 labeled pixels out of a total of 318 023). Therefore, the active learning process is naturally much more difficult. To obtain convergence, a reestimation of the parameters has been done at iteration 30 (when the solution stabilized for the three active methods to a suboptimal result).

The Full SVM achieves an error of 9.77% on the test set, with a kappa index of 0.870. Results for the active learning algorithms are shown in Fig. 7(b) for the learning curves and in Fig. 9 for the classification maps. The curves in Fig. 7(b) show that only the EQB is able to converge to the Full SVM test error. This is due mainly to the parameter reestimation at iteration 30 (about 3400 pixels) that allows the method to converge to the true minimum. The speed of the convergence of EQB is similar to the one observed for the other active methods, confirming its efficiency.

On the contrary, MS and MS-cSV do not improve with the reestimation of the parameters, and their results equal the ones obtained without reestimation. This is due to the fact that these

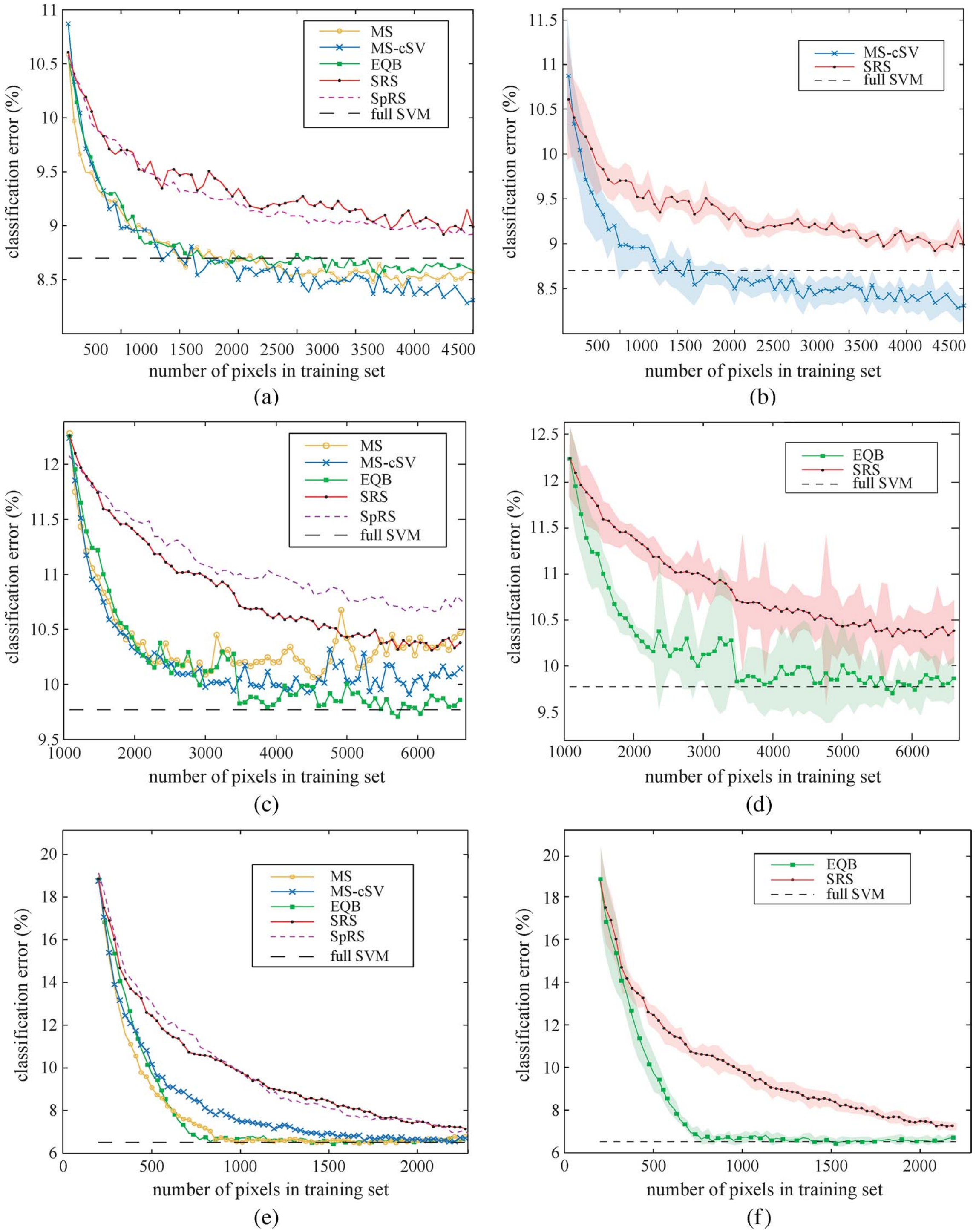


Fig. 7. Classification error curves for (a) Rome, (c) Las Vegas, and (e) KSC. Each curve shows the mean error for growing size training sets over several runs of the algorithm starting with different initial sets X. (d), (e), and (f) show the error bars of the best model against the SRS. The shaded areas show the standard deviation of over the results of the independent runs considered. MS = margin sampling, MS-cSV = margin sampling by closest support vectors, EQB = entropy query-by-bagging, and SpRS = spatial random sampling.

methods depend on the margin of the SVM, which is modified at every update of X. After every update, the current margin is only refined. However, when reestimating the kernel parameters Θ, the margin changes radically, presenting to the algorithms a new active learning setting.

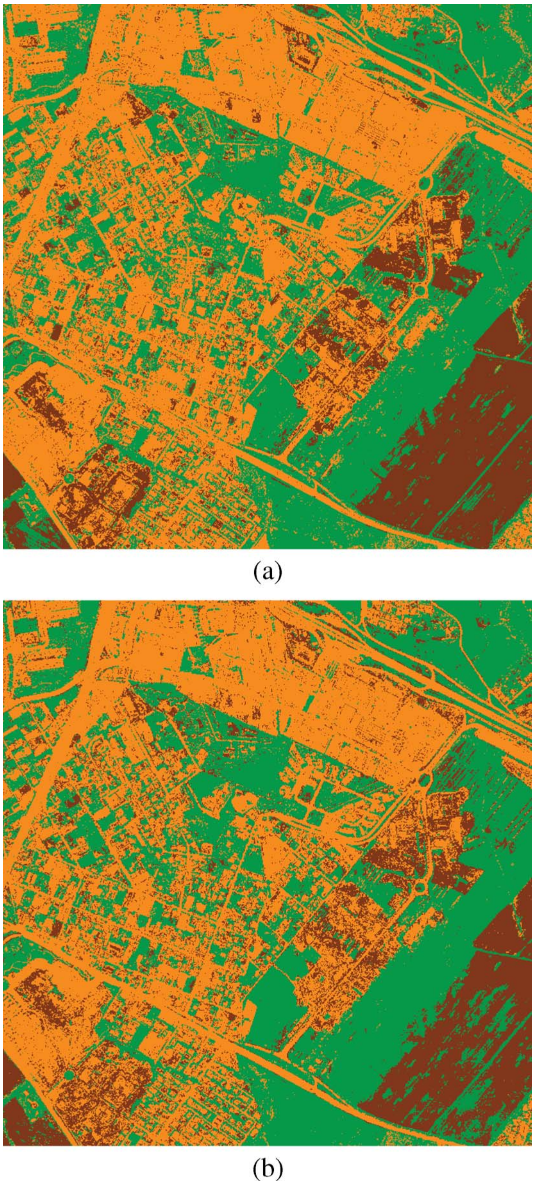
Despite the nonconvergence, we can observe that the MS-cSV performs better that the MS algorithm, taking advantage of the distribution of the pixels in Q after the first iterations, where both methods perform similarly. Moreover, the MS method cannot converge and is equivalent to SRS after the 50th iteration. An intuitive explanation is that MS-cSV avoids oversampling in dense regions close to the margin and samples all the feature space equivalently.

Regarding global performances at iteration 31 (see Table VI), EQB shows the best results both in terms of accuracy (89.78%) and kappa statistics (0.866): these results corresponded to the

TABLE V

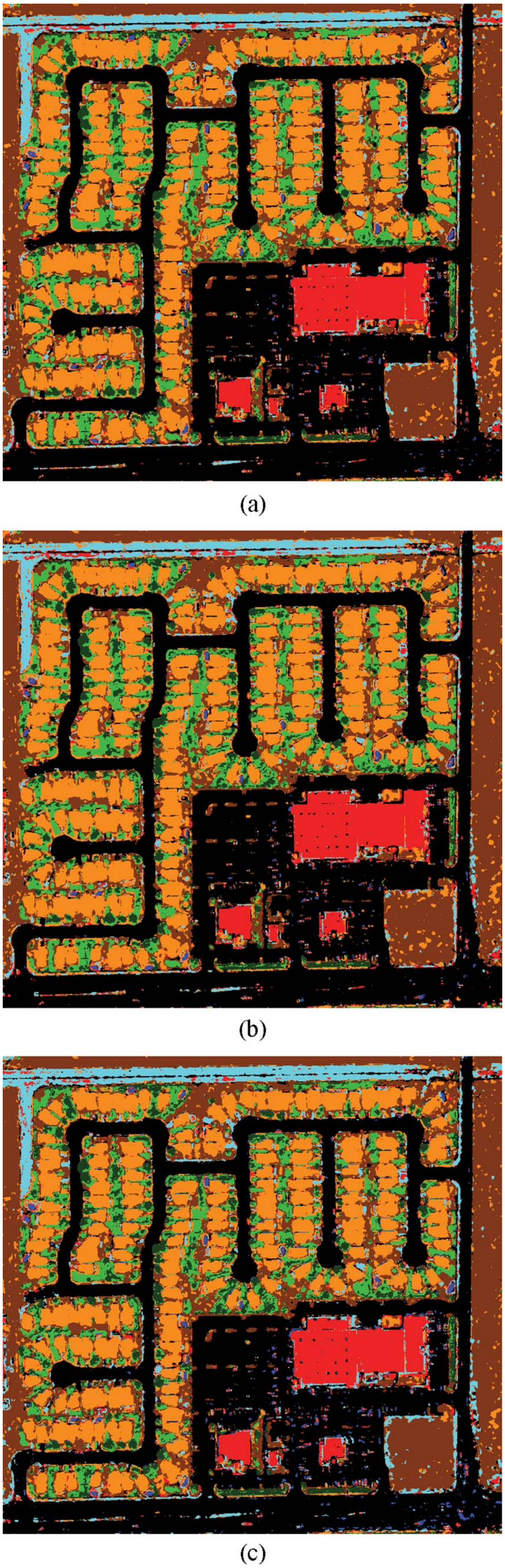
CLASS ACCURACIES (%) AND K INDEX FOR THE ROME DATA SET. STD = STANDARD DEVIATION, CONF. = CONFIDENCE INTERVAL, AND ∗ = SIGNIFICANTLY DIFFERENT FROM SRS (Z TEST, [50])

		Full	MS	MS-cSV	EQB	SRS
iteration		-	26	26	26	26
training pixels		18000	1860	1860	1860	1860
Man made		93.44	**94.21**	94.12	**94.21**	93.85
Vegetation		93.77	**91.44**	91.03	90.77	90.28
Soil		83.73	81.90	**82.93**	81.61	80.46
Overall accuracy		91.30	91.30	**91.43**	91.19	90.64
K	Mean	0.810	0.809*	**0.813***	0.807*	0.794
	Std.		0.0033	0.0025	0.0042	0.0049
	Conf. (α = 5%)		[0.807; 0.811]	[0.812; 0.814]	[0.804; 0.808]	[0.791; 0.797]


(a)

(b)

Fig. 8. Classification map of the Rome image using (a) Full SVM and (b) MS-cSV (orange = man made, green = vegetation, and brown = soil).


(a)

(b)

(c)

Fig. 9. Classification maps of the Las Vegas image using (a) Full SVM, (b) EQB, and (c) SRS (orange = residential buildings, red = commercial buildings, black = asphalt, light green = short vegetation, dark green = trees, brown = soil, blue = water, and cyan = drainage channel).

good performance for the main classes of the image, where EQB even outperforms the full SVM (residential and commercial buildings). MS-cSV also shows good performances

TABLE VI
CLASS ACCURACIES (%) AND K INDEX FOR THE LAS VEGAS DATA SET. STD = STANDARD DEVIATION, CONF. = CONFIDENCE INTERVAL, AND $*$ = SIGNIFICANTLY DIFFERENT FROM SRS (Z TEST, [50])

		Full	MS	MS-cSV	EQB	SRS
iteration		-	31	31	31	31
training pixels		30000	3480	3480	3480	3480
Residential b.		85.24	84.75	84.86	**85.55**	82.25
Commercial b.		84.33	82.79	82.87	**84.77**	82.88
Asphalt		98.31	**98.23**	98.19	97.90	98.00
Short veg.		84.55	**82.59**	82.25	81.99	78.02
Trees		58.37	59.38	60.50	59.53	**63.61**
Soil		92.20	91.71	91.79	91.14	**91.88**
Water		67.35	65.86	65.10	65.07	**72.37**
Drainage ch.		80.47	77.21	78.27	79.56	**81.93**
Overall accuracy		90.23	89.64	89.73	**89.78**	89.09
K	Mean	0.870	0.863	0.864*	**0.866***	0.855
	Std.		0.0054	0.0030	0.0020	0.0030
	Conf.		[0.859;	[0.862;	[0.865;	[0.853;
	(α = 95%)		0.867]	0.866]	0.867]	0.857]

(accuracy = 89.73%, kappa = 0.864), higher than MS results for five out of eight classes.

Looking at the accuracies per class (Table VI), SRS shows the best performance for the classes "Trees," "Drainage channel," and (in particular) "Water." These classes are the most scarce in the ground survey. This result can be explained by the VHR of the image: for the active methods [see Fig. 9(b)], the main sources of errors are due to small objects such as cars, chimneys, road lines, or dry bushes that contaminate the spectral signature of the main class. This can degrade the performances of an active learner. For instance, cars do not have a specific class and are included in the ground survey in the class "Asphalt." An SVM trained on spectral values only will have the tendency to misinterpret these pixels and classify them as water or soil.

For an active learner, this kind of pixel has a high probability to be included in the training set because they are contradictory with respect to the class indicated by the ground survey. This causes a displacement of the decision boundary between "Asphalt" and "Water"/"Soil" into a zone otherwise clear. On the one hand, such a displacement results in the improvement of the result for the class "Asphalt" which becomes more robust to noise caused by small objects. However, on the other hand, accuracies for classes "Water" or "Soil" are degraded, because spectral responses typical to these classes are classified as "Asphalt."

On the contrary, the SRS ignores these uncommon pixels: most of the pixels being unmixed in the classes "Asphalt" and "Residential buildings," a random selection naturally pays little attention to pixels related to small objects. However, kappa statistics take into account the higher commission errors. The kappa index (Table VI) and the mappings in Fig. 9(a)–(c) confirm this hypothesis. Small objects are labeled as "Water" and "Soil" by SRS much more than by the other two mappings: even if water or soil pixels of the ground survey are better classified by SRS, commission errors remain important for the classes "Asphalt" and "Residential buildings" [see Fig. 9(c)].

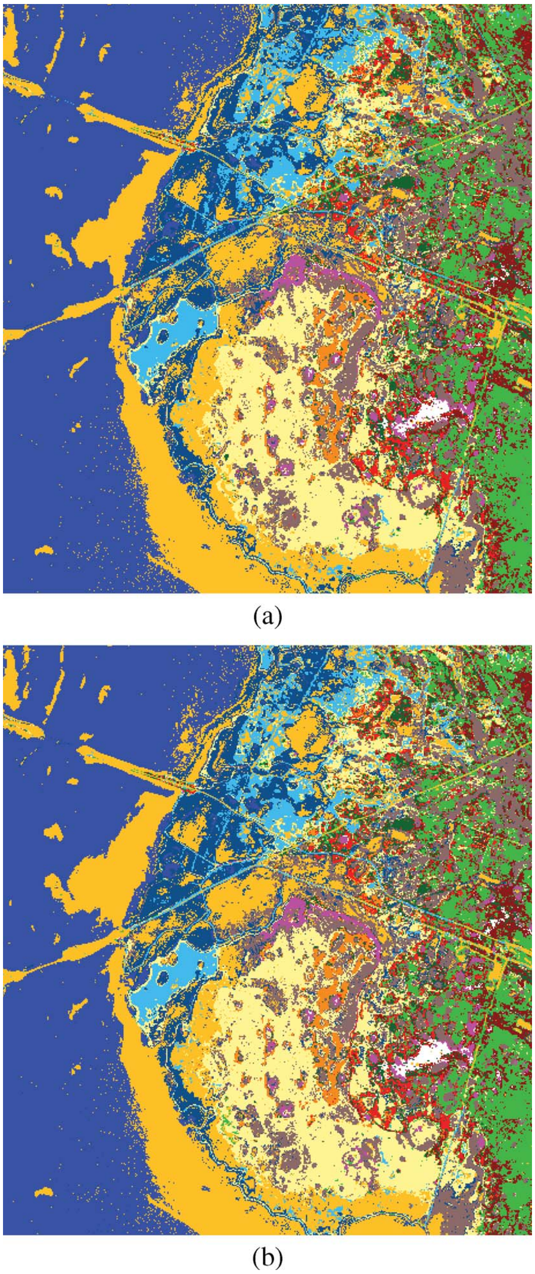
These considerations raise the question of the resolution required for a classification task. In this case, the resolution is so high that objects introduce noise and degrade the solution.

C. KSC

For the KSC image, the Full SVM attains a test error of 6.52%—equaling the accuracy achieved in [46]—and a kappa index of 0.928. All the active learning algorithms converge to the lower bound at different speeds: the faster convergence is met by the EQB algorithm that reaches the Full SVM error in about 20 iterations, i.e., with a training set of 800 pixels. MS and MS-cSV show a faster convergence in the first iterations. EQB shows a constant decrease and shows the best results in terms of overall accuracy (93.36%): this is related to the excellent results obtained for the classes "Scrub" and "Water," the most represented in the test set. MS gives the best results in most of the classes and in terms of kappa coefficient (0.919). Therefore, MS and EQB models seem to be the most appropriate for this data set. MS-cSV converges to the Full SVM slower than the two other methods, as shown in Fig. 10(c). That can be explained by the results in Table VII: MS-cSV cannot find the optimal solution for the mixed classes, such as classes "Cabbage palm/hammock" and "Oak/broadleaf hammock" (in italic on Table VII). These classes are very close in the hyperspectral space and can impair the choice of the closest support vectors: mixed classes result in pixels lying in the same regions of the feature space but belonging to different classes. Since the MS-cSV avoids picking several training points in the same dense area, the constraint on density of the candidates avoids picking them simultaneously, despite their importance. These pixels are sampled slower than that with the MS algorithm, resulting in slower convergence to the Full SVM by the smaller accuracies on mixed classes. Therefore, MS-cSV seems to be less efficient in the presence of overlapping classes. Nonetheless, for nonmixed classes, MS-cSV often gives the best result in terms of overall accuracy.

D. Robustness to Ill-Posed Scenarios

In an ill-posed scenario, where only a limited amount of labeled pixels per class is available in the initial training set X, the model built at the first iteration could fail to represent the true data distribution. Then, there is a risk that the candidates selected were not the most relevant to decrease the classification error. This could be particularly important for the EQB algorithm, where the entropy is computed over a committee of suboptimal classifiers. In this case, the selection is done in the wrong region of the feature space, and there is no reason to believe that it would be worst than SRS. However, the benefits of EQB appear after a few iterations, as soon as a sufficient amount of pixels is selected to train the k models of the committee. Fig. 11 shows this principle for the KSC image and starting with one labeled pixel per class (starting size of X is 13 pixels) and adding 30 pixels per iteration: after four iterations, the EQB algorithm starts to outperform SRS and converges to the Full SVM result when using about 800 pixels, equaling the results reported in Table VII.



(a)

(b)

Fig. 10. Classification of the KSC image using (a) Full SVM and (b) EQB (see Table IV for legend colors).

V. Conclusion

In this paper, we have presented an active learning framework for remote sensing image classification. In this framework, the predictor has control on the composition of the training set and chooses the most valuable pixels for the improvement of its performance. A state-of-the-art active learning method, the SVM-based MS [15], has been discussed, and two novel methods have been presented and applied for the classification of VHR urban scenes.

The first method, MS-cSV, is a novel modification of the MS that takes the distribution of the unlabeled candidates in the feature space into account: this way, oversampling on dense regions is avoided, as is the risk of not sampling important regions.

TABLE VII
Class Accuracies (%) and K Index for the KSC Data Set. STD = Standard Deviation, Conf. = Confidence Interval, and $*$ = Significantly Different From SRS (Z Test, [50])

		Full	MS	MS-cSV	EQB	SRS
iteration		-	20	20	20	20
training pixels		2500	800	800	800	800
Scrub		94.79	94.66	94.60	**94.79**	93.23
Willow swamp		88.14	**85.59**	85.38	84.53	81.36
Cabbage palm hammock		86.96	**86.59**	83.51	86.05	82.61
Cabbage palm/oak hammock		*75.00*	***76.73***	*58.85*	*68.46*	*55.39*
Slash pine		91.18	**91.07**	83.21	87.86	75.71
Oak/b. hammock		*75.56*	***76.73***	*65.77*	*69.42*	*57.50*
Hardwood swamp		83.78	**87.50**	84.87	81.25	82.57
Graminoid marsh		94.17	84.94	**94.31**	92.97	92.67
Spartina marsh		97.96	**97.70**	97.36	96.94	95.58
Cattail marsh		98.26	96.30	**98.04**	96.41	94.89
Salt marsh		99.11	98.66	**99.11**	97.43	96.87
Mud flats		93.06	88.80	**92.71**	91.41	90.28
Water		98.83	98.15	**98.74**	98.69	98.59
Overall accuracy		93.49	93.32	92.15	**93.36**	89.39
K	Mean	0.928	**0.919***	0.907*	0.910*	0.882
	Std.		0.0047	0.0057	0.0058	0.0054
	Conf. $\alpha = 95\%$		[0.915 ; 0.923]	[0.902 ; 0.912]	[0.905 ; 0.915]	[0.877 ; 0.887]

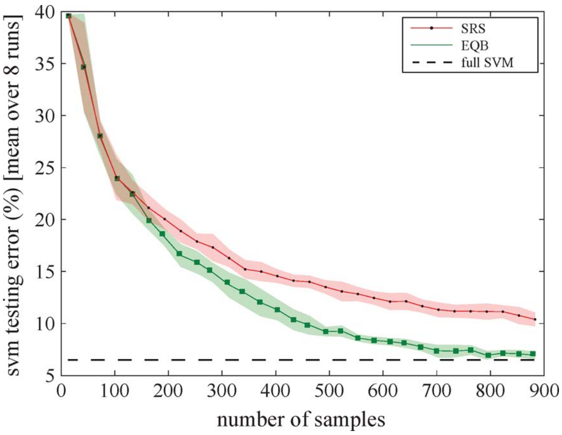


Fig. 11. Results of EQB and SRS for the KSC image in an ill-posed scenario, where only 1 pixel per class is considered in X. Initial size of X is 13 pixels, and 30 pixels are added at each iteration (markers on the curves). Shaded regions show the standard deviation of the predictions obtained over eight independent runs of the algorithms.

The second method, EQB, is independent of the classifier and is based on committee learning: a committee of predictors labels the candidates, and the entropy in the distribution of the predictions of the candidates is used as heuristic.

Applications on VHR QuickBird images and on AVIRIS hyperspectral images showed the consistency of the methods. Training sets created actively can perform as good as a predictor trained on a complete ground survey (Full SVM). Table VIII resumes the main results for the three images considered. For actively selected training sets, about 10% of the full SVM

TABLE VIII
KAPPA INDEX FOR THE DATA SETS CONSIDERED

Method	Rome 2.4 m	Las Vegas 0.6 m	KSC 18 m
Full SVM	0.810	0.870	0.928
MS	0.809	0.863	**0.919**
MS-cSV	**0.813**	0.864	0.907
EQB	0.806	**0.866**	0.910
SRS	0.794	0.855	0.882

size is necessary to converge to the same results in terms of both classification accuracy and mapping of the whole scene. For all the applications considered, the convergence of the methods proposed to the optimal result is quicker than the SRS, confirming the interest of active learning methods. In particular, EQB has shown excellent performances for all the data sets considered. The performances of the method are at least comparable to that of the MS, which is optimal for SVMs. The novelty of the EQB method, which lies in its independence to the classifier used, opens new possibilities of application for the method.

MS-cSV provides an interesting update of the classical MS method in order to handle inclusion of several pixels at each iteration. The method has shown better performances than MS on the QuickBird case studies, improving the MS efficiency with the same convergence speed. Nonetheless, the method still needs improvement in order to handle situations with mixed classes, where the constraint on the closest support vector slows the speed of convergence.

Issues related to too very small objects such as cars in VHR imagery and the problems that are raised by their inclusion in the training set have been addressed. Active methods, as well as the models run on the whole ground survey (Full SVM), suffer from this problem. Nonetheless, both the methods proposed showed enough robustness to result in very high accuracies, particularly for the main classes of the images considered, and the more often show higher accuracies than the MS.

Finally, all the active learning methods depend heavily on the quality of the initial data: the initial training set being very small, there is a risk that a part of the feature space is not covered. If the uncovered part is in an area that the current predictor considers as correctly handled, it will be impossible to sample points from that area. Further development will aim at making the algorithms less deterministic, i.e., allowing them to choose candidates with a certain probability proportional to the heuristic considered instead of only selecting the pixels related to maximum entropy/minimum distance. Moreover, parameter estimation could be updated during the process, for instance, by using optimization algorithms to adjust the parameters during the active learning algorithm, avoiding successive grid search procedures.

APPENDIX A
PSEUDOCODE OF THE ACTIVE LEARNING ALGORITHMS CONSIDERED

Algorithm 1 Margin sampling (MS)

```
Inputs
 —Initial training set $X$.
 —Set of candidates $Q$.
 —Number of classes $(Ncl)$.
 —Pixels to add at every iteration $(Npts)$.
  for each iteration do
    Train current classifier with current training set $X$.
    Compute test error of the current classifier.
    for each class $cl$ do
      for each candidate $q_i$ to add do
        Compute the distance to the margin for the candidate $q_i$ for class $cl$ using (1). The result is a $(m \times Ncl)$ distance matrix.
      end for
    end for
    Compute the minimum distance over the $Ncl$ classes. The result is a $(m \times 1)$ distance vector.
    Label the $Npts$ pixels associated with minimum distance.
    Update $X$ with the $Npts$ chosen pixels and remove those from $Q$.
  end for
```

Algorithm 2 Margin sampling by cSV (MS-cSV)

```
Inputs
 —Initial training set $X$.
 —Set of candidates $Q$.
 —Number of classes $(Ncl)$.
 —Pixels to add at every iteration $(Npts)$.
  for each iteration do
    Train current classifier with current training set $X$.
    Compute test error of the current classifier.
    for each class $cl$ do
      for each candidate $q_i$ to add do
        Compute the distance to the margin for the candidate $q_i$ for class $cl$ using (1). The result is a $(m \times Ncl)$ distance matrix.
        Select the support vector $j$ that minimizes $K(x_j, q_i)$. The result is a $(m \times Ncl)$ support vector list $(cSV)$.
      end for
    end for
    Compute minimal distance over the $Ncl$ classes. The result is a $(m \times 1)$ distance vector.
    Add $q_1$ to a temporary list $G$.
    for $i = 2$ to $m$ do
      if $cSV_i \neq cSV_{i-1}$ then
        Add the candidate related to $\min_{q \in G} |f(q_i)|$ to the best candidate list $B$.
        Clear $G$.
        Add $q_i$ to a temporary list $G$.
      else
        Add $q_i$ to a temporary list $G$.
      end if
    end for
    Label the $Npts$ pixels associated with minimal distance in $B$.
    Update $X$ with the $Npts$ chosen pixels. Remove the selected pixels from $Q$.
  end for
```

```
Algorithm 3 Entropy-based query-by-bagging (EQB)
Inputs
  —Initial training set $X$.
  —Set of candidates $Q$.
  —Pixels to add at every iteration ($Npts$).
  —Number of bootstrap samples ($k$).
  —Share of $X$ drawn into the bootstrap samples ($pct$).
    for each iteration do
      Train current classifier with current training set $X$.
      Compute test error of the current classifier.
      for $t = 1$ to $k$ do
        By resampling according to $U(x)$ on $X$, obtain subset $X'_t$ of size $pct * X$.
        Train the $t$-th SVM on $X'_t$.
        Predict the class membership $f_t(q_i)$ of the $m$ candidates $\in Q$. The result is a $(m \times k)$ vector.
      end for
      Compute the entropy $H(q_i)$ for every candidate.
      Label the $Npts$ pixels associated with maximum entropy.
      Update $X$ with the $Npts$ chosen pixels. Remove the $Npts$ pixels from $Q$.
    end for
```

Acknowledgment

The authors would like to thank DigitalGlobe for providing data which have been used in this paper and the authors in [46] for the KSC data.

References

[1] B. Boser, I. Guyon, and V. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proc. 5th ACM Workshop Comput. Learn. Theory*, 1992, pp. 144–152.

[2] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[3] C. Huang, L. S. Davis, and J. R. G. Townshend, “An assessment of support vector machines for land cover classification,” *Int. J. Remote Sens.*, vol. 23, no. 4, pp. 725–749, 2002.

[4] G. M. Foody and A. Mathur, “A relative evaluation of multiclass image classification by support vector machines,” *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 6, pp. 1335–1343, Jun. 2004.

[5] J. Inglada, “Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features,” *ISPRS J. Photogramm. Remote Sens.*, vol. 62, no. 3, pp. 236–248, Aug. 2007.

[6] 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.

[7] G. Camps-Valls, L. Gómez-Chova, J. Calpe, E. Soria, J. D. Martín, L. Alonso, and J. Moreno, “Robust support vector method for hyperspectral data classification and knowledge discovery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 7, pp. 1530–1542, Jul. 2004.

[8] G. Camps-Valls and L. Bruzzone, “Kernel-based methods for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.

[9] M. Fauvel, J. Chanussot, and J. A. Benediktsson, “Evaluation of kernels for multiclass classification of hyperspectral remote sensing data,” in *Proc. IEEE ICASSP*, Toulouse, France, 2006, pp. II-813–II-816.

[10] M. Chi, R. Feng, and L. Bruzzone, “Classification of hyperspectral remote-sensing data with primal SVM for small-sized training dataset problem,” *Adv. Space Res.*, vol. 41, no. 11, pp. 1793–1799, 2008.

[11] V. Castelli and T. M. Cover, “On the exponential value of labeled samples,” *Pattern Recognit. Lett.*, vol. 16, no. 1, pp. 105–111, Jan. 1995.

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

[13] D. Cohn, L. Atlas, and R. Ladner, “Improving generalization with active learning,” *Mach. Learn.*, vol. 15, no. 2, pp. 201–221, May 1994.

[14] D. Cohn, Z. Ghahramani, and M. I. Jordan, “Active learning with statistical models,” *J. Artif. Intell. Res.*, vol. 4, pp. 129–145, 1996.

[15] G. Schohn and D. Cohn, “Less is more: Active learning with support vectors machines,” in *Proc. 17th ICML*, Stanford, CA, 2000, pp. 839–846.

[16] C. Campbell, N. Cristianini, and A. Smola, “Query learning with large margin classifiers,” in *Proc. ICML*, Stanford, CA, 2000, pp. 111–118.

[17] H. T. Nguyen and A. Smeulders, “Active learning using pre-clustering,” in *Proc. 21th ICML*, Banff, AB, Canada, 2004, p. 79.

[18] A. Pozdnoukhov and M. Kanevski, “Monitoring network optimisation for spatial data classification using support vector machines,” *Int. J. Environ. Pollut.*, vol. 28, no. 3/4, pp. 465–484, 2006.

[19] 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, pp. 589–613, Dec. 2005.

[20] X. Li, L. Wang, and E. Sung, “Multilabel SVM active learning for image classification,” in *Proc. ICIP*, Singapore, 2004, pp. 2207–2210.

[21] F. Jing, M. Li, H. Zhang, and B. Zhang, “Entropy-based active learning with support vector machines for content-based image retrieval,” in *Proc. IEEE ICME*, Taipei, Taiwan, 2004, pp. 85–88.

[22] S. Cheng and F. Y. Shih, “An improved incremental training algorithm for support vector machines using active query,” *Pattern Recognit.*, vol. 40, no. 3, pp. 964–971, Mar. 2007.

[23] P. H. Gosselin and M. Cord, “Precision-oriented active selection for interactive image retrieval,” in *Proc. IEEE ICIP*, Atlanta, GA, 2006, pp. 3197–3200.

[24] M. Ferecatu and N. Boujemaa, “Interactive remote-sensing image retrieval using active relevance feedback,” *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 818–826, Apr. 2007.

[25] 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, Mar. 2002.

[26] C. Silva and B. Ribeiro, “Margin-based active learning and background knowledge in text mining,” in *Proc. 4th Int. Conf. HIS*, Washington, DC, 2004, pp. 8–13.

[27] P. Mitra, C. A. Murphy, 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.

[28] D. D. Lewis and W. A. Gale, “A sequential algorithm for training text classifiers,” in *Proc. 17th Annu. Int. ACM-SIGIR Conf. Res. Dev. Inf. Retrieval*, W. B. Croft and C. J. van Rijsbergen, Eds., London, U.K., pp. 3–12.

[29] N. Roy and A. McCallum, “Toward optimal active learning through sampling estimation of error reduction,” in *Proc. ICML*, Williamstown, MA, 2001, pp. 441–448.

[30] S. Kullback and R. A. Leibler, “On information and sufficiency,” *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[31] H. S. Seung, M. Opper, and H. Sompolinski, “Query by committee,” in *Proc. Annu. Workshop Comput. Learn. Theory*, New York, 1992, pp. 287–294.

[32] Y. Freund, H. Seung, E. Shamir, and N. Tishby, “Selective sampling using the query by committee algorithm,” *Mach. Learn.*, vol. 28, no. 2/3, pp. 133–168, Aug. 1997.

[33] I. Dagan and S. P. Engelson, “Committee-based sampling for training probabilistic classifiers,” in *Proc. ICML*, San Francisco, CA, 1995, pp. 150–157.

[34] P. Melville, “Creating diverse ensemble classifiers to reduce supervision,” Ph.D. dissertation, Univ. Texas, Austin, Austin, TX, 2005.

[35] L. I. Kuncheva, *Combining Pattern Classifiers*. Hoboken, NJ: Wiley-Interscience, 2004.

[36] N. Abe and H. Mamitsuka, “Query learning strategies using boosting and bagging,” in *Proc. ICML*, Madison, WI, 1998, pp. 1–9.

[37] Y. Freund and R. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1999.

[38] L. Breiman, “Bagging predictors,” Univ. California, Berkeley, Berkeley, CA, Tech. Rep. 421, 1994.

[39] P. Melville and R. J. Mooney, “Diverse ensembles for active learning,” in *Proc. ICML*, Banff, AB, Canada, 2004, p. 74.

[40] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, “Text classification from labeled and unlabeled documents using EM,” *Mach. Learn.*, vol. 39, no. 2/3, pp. 103–134, May/Jun. 2000.

[41] 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.

[42] 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.

[43] G. Jun and J. Ghosh, "An efficient active learning algorithm with knowledge transfer for hyperspectral remote sensing data," in *Proc. IEEE IGARSS*, Boston, MA, 2008.

[44] Y. Zhang, X. Liao, and L. Carin, "Detection of buried targets via active selection of labeled data: Application to sensing subsurface UXO," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 11, pp. 2535–2543, Nov. 2004.

[45] Q. Liu, X. Liao, and L. Carin, "Detection of unexploded ordnance via efficient semisupervised and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 9, pp. 2558–2567, Sep. 2008.

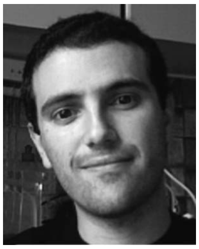
[46] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2008.

[47] M. Chini, F. Pacifici, W. J. Emery, N. Pierdicca, and F. Del Frate, "Comparing statistical and neural network methods applied to very high resolution satellite images showing changes in man-made structures at rocky flats," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1812–1821, Jun. 2008.

[48] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Stat.*, vol. 7, no. 1, pp. 1–26, 1979.

[49] R. Collobert, S. Bengio, and J. Mariéthoz, "Torch: A modular machine learning software library," IDIAP, Martigny, Switzerland, Tech. Rep. IDIAP-RR 02-46, 2002.

[50] G. M. Foody, "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *Photogramm. Eng. Remote Sens.*, vol. 50, no. 5, pp. 627–633, 2004.



Devis Tuia (S'07) received the Diploma in geography from the University of Lausanne, Lausanne, Switzerland, in 2004, and the Master of Advanced Studies in environmental engineering from the Federal Institute of Technology, Lausanne, Switzerland, in 2005. He is currently working toward the Ph.D. degree in the field of machine learning and its applications to urban remote sensing at the Institute of Geomatics and Analysis of Risk, University of Lausanne.

Mr. Tuia was one of the winners of the 2008 IEEE GEOSCIENCE AND REMOTE SENSING Data Fusion Contest.

Frédéric Ratle received the degree in engineering physics and the Master of Applied Sciences in the field of optimization and statistical methods in mechanical engineering from the Ecole Polytechnique de Montreal, Montreal, QC, Canada, in 2003 and 2005. He is currently working toward the Ph.D. degree in machine learning and data analysis at the Institute of Geomatics and Analysis of Risk, University of Lausanne, Lausanne, Switzerland.

Mr. Ratle was one of the winners of the 2008 IEEE GEOSCIENCE AND REMOTE SENSING Data Fusion Contest.

Fabio Pacifici (S'03) was born in Rome, Italy, in 1980. He received the B.S. (*cum laude*) and M.S. (*cum laude*) degrees in telecommunication engineering from "Tor Vergata" University, Rome, in 2003 and 2006, respectively. He is currently working toward the Ph.D. degree in geoinformation at the Earth Observation Laboratory, "Tor Vergata" University.

Since 2005, he has been collaborating with the Department of Aerospace Engineering Sciences, University of Colorado, Boulder. He is currently involved in various remote-sensing projects supported by the European Space Agency and the Italian Space Agency, with emphasis on neural network applications. His research interests include remote-sensing image processing, analysis of multitemporal imagery, and data fusion, particularly classification and change detection of urban areas using very high-resolution optical and synthetic aperture radar imagery.

Mr. Pacifici is a Reviewer for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He was the recipient of the First Prize in both the 2007 and the 2008 IEEE GEOSCIENCE AND REMOTE SENSING Data Fusion Contest.

Mikhail F. Kanevski received the Ph.D. degree in plasma physics from the Moscow State University, Moscow, Russia, in 1984, and the Doctoral theses in computer science from the Institute of Nuclear Safety, Russian Academy of Science, Moscow, in 1996.

Until 2000, he was a Professor with the Moscow Physico–Technical Institute (Technical University) and the Head of laboratory with the Moscow Institute of Nuclear Safety, Russian Academy of Sciences. Between 1999 and 2002, he was an Invited Professor with the IDIAP Research Institute, Switzerland. Since 2004, he has been a Professor with the Institute of Geomatics and Analysis of Risk, University of Lausanne, Lausanne, Switzerland. He is a Principal Investigator of several national and international grants. His research interests include geostatistics for spatio-temporal data analysis, environmental modeling, computer science, numerical simulations, and machine learning algorithms. Remote sensing image classification, natural hazard assessments (forest fires, avalanches, and landslides), and time series predictions are the main applications considered at his laboratory.

William J. Emery (M'90–SM'01–F'02) received the Ph.D. degree in physical oceanography from the University of Hawaii, Mãnoa, in 1975.

After being with Texas A&M University, College Station, he was with the University of British Columbia, Vancouver, BC, Canada, in 1978, where he created a satellite oceanography facility and education/research program. Since 1987, he has been a Full Professor with the Department of Aerospace Engineering Sciences, University of Colorado, Boulder. He is active in the analysis of satellite data for oceanography, meteorology, and terrestrial physics (vegetation, forest fires, sea ice, etc.). His research focus areas are in satellite sensing of sea surface temperature, mapping ocean surface currents (imagery and altimetry), sea ice characteristics/motion, and terrestrial surface processes. He has recently started working in urban change detection using high-resolution optical imagery and synthetic aperture radar data. This is done with students from various universities in Rome, Italy, where he is an Adjunct Professor in geoinformation with the "Tor Vergata" University, Rome. He also works with passive microwave data for polar applications to ice motion and ice concentration and to atmospheric water vapor studies. In addition, his group writes image navigation and analysis software and has established/operated data systems for the distribution of satellite data received by their own antennas. He is a coauthor of two textbooks on physical oceanography, has translated three oceanographic books (German to English), and the author of over 130 published articles.

Dr. Emery is a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society and the Past Editor of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He is an associate member of the Laboratory for Atmospheric and Space Physics, an affiliate member of NOAA's Cooperative Institute for Research in Earth Science, and a founding member of the Program in Oceanic and Atmospheric Sciences, which is now the Department of Atmospheric and Ocean Sciences.