Title: Definition of Effective Training Sets for Supervised Classification of Remote Sensing Images by a Novel Cost-Sensitive Active Learning Method

This paper appears in: IEEE Transactions on Geoscience and Remote Sensing

Date of Publication: 2014

Author(s): Begüm Demir, Luca Minello, Lorenzo Bruzzone,

Volume:52, Issue: 2

Page(s): 1272 - 1284

DOI: 10.1109/TGRS.2013.2249522

Definition of Effective Training Sets for Supervised Classification of Remote Sensing Images by a Novel Cost-Sensitive Active Learning Method

Begüm Demir, *Member IEEE*, Luca Minello, *and* Lorenzo Bruzzone, *Fellow, IEEE*

Dept. of Information Engineering and Computer Science, University of Trento,
Via Sommarive, 14, I-38123 Trento, Italy
e-mail: demir@disi.unitn.it, luca.minello@disi.unitn.it, lorenzo.bruzzone@ing.unitn.it.

***Abstract*—** This paper proposes a novel cost-sensitive active learning (CSAL) method to the definition of reliable training sets for the classification of remote sensing images with Support Vector Machines. Unlike the standard active learning (AL) methods, the proposed CSAL method redefines AL by assuming that the labeling cost of samples during ground survey is not identical, but depends on both the samples accessibility and the traveling time to the considered locations. The proposed CSAL method selects the most informative samples on the basis of three criteria: i) uncertainty, ii) diversity, and iii) labeling cost. The labeling cost of the samples is modeled by a novel cost function that exploits ancillary data such as the road network map and the digital elevation model of the considered area. In the proposed method, the three criteria are applied in two consecutive steps: in the first step the most uncertain samples are selected, whereas in the second step the uncertain samples that are diverse and have low labeling cost are chosen. In order to select the uncertain samples that optimize the diversity and cost criteria, we propose two different optimization algorithms. The first algorithm is defined on the basis of a sequential forward selection optimization strategy, whereas the second one relies on a genetic algorithm. Experimental results show the effectiveness of the proposed CSAL method compared to standard

active learning methods that neglect the labeling cost.

Index Terms – Active learning, automatic classification, ground data collection, training set, genetic algorithm, sequential forward selection, remote sensing.

I. INTRODUCTION

Generation of land-cover maps is one of the most common applications in remote sensing image analysis and is usually achieved by supervised classification techniques. Such techniques require the availability of reliable ground reference samples to be used in the learning phase of the classification algorithm. Reliability of the labeled training samples depends on both the quantity and quality of the available samples. In remote sensing, the quality of samples is affected by two main issues [1]: a) the capability to model the spatial variability of the spectral signatures of the land-cover classes, and b) the high spatial correlation among the training samples collected in neighboring locations of the same area (that reduces the information conveyed by training samples with respect to the case of independent samples) [1]. Thus, the collection of a sufficient number of reliable labeled samples is time consuming, complex and costly in operational scenarios, and can significantly affect the final accuracy of produced land-cover maps.

To overcome the above-mentioned problems, active learning (AL) methods have been recently presented in the remote sensing literature. They aim to optimize the definition of a training set by selecting a minimum number of high quality labeled samples on the basis of an iterative process [2]-[11]. In other words, AL avoids the collection of labels for non-useful (i.e., redundant) samples. This also leads: i) to reduce the computational complexity for the learning phase (thanks to the optimized training set with a small number of labeled samples), and ii) to increase the classification accuracy (thanks to the improved class models estimated on a high quality training set defined on the basis of the classification rule of the considered classifier).

Most of the AL works presented in the remote sensing literature select a batch of informative samples at each iteration on the basis of two main criteria: 1) uncertainty, and 2) diversity of samples. The uncertainty criterion is associated to the confidence of the supervised algorithm in correctly classifying the considered sample, whereas the diversity criterion aims at selecting a set of unlabeled samples that are as more diverse as possible to reduce the redundancy among them [8]. An important drawback of most of the above-mentioned AL methods is the assumption that the cost of labeling samples is equal for all the samples on the ground. This assumption may be not appropriate for remote sensing problems. For these reasons, AL methods presented in the remote sensing literature do not fit well with the real applications constraints by assuming an uniform labeling cost for all the samples on the ground.

In remote sensing, the labeling of the selected samples during AL process can be carried out by i) *in situ* ground surveys (which are associated to high cost), ii) image photointerpretation (which is cheaper, yet can be applied only in limited cases), or iii) hybrid strategies (integration of photointerpretation and ground surveys). It is worth noting that ground surveys are mandatory when detailed classes should be recognized. The cost for the labeling process in this case is generally high as a) traveling to locations of the selected samples is required to identify the land-cover types of the related area on the ground, and b) some geographical areas may be not easily reachable in operational scenarios. Thus, the label acquisition cost for each sample *in situ* ground surveys depends on the accessibility of the sample and also on the traveling time to reach the related locations.

On the basis of these analyses, we can state that the assumption that all samples have the same labeling cost often does not hold in remote sensing problems, when the labeling of the samples is carried out by in situ ground surveys. In order to consider the labeling cost of samples during the AL process, in this paper we present a novel cost-sensitive AL (CSAL) method to the classification of remote sensing images by a Support Vector Machine (SVM) classifier. Differently

from the AL methods presented in the remote sensing literature, the proposed method is defined on the basis of the evaluation of three criteria: i) uncertainty, ii) diversity and iii) labeling cost. The proposed method defines the labeling cost with respect to both i) samples accessibility and ii) traveling time required to visit the selected samples. The accessibility of the samples and the traveling time between the samples are modeled by using ancillary data such as the road network map and the digital elevation model (DEM) of the considered area. The proposed method assesses the uncertainty of samples by the Multiclass Level Uncertainty (MCLU) technique presented in [8] (which proved to be effective for multiclass classification problems), whereas jointly evaluates the diversity and cost of the samples by two different algorithms. The first algorithm is defined on the basis of a simple sequential forward selection strategy, whereas the second one is based on a genetic algorithm.

The paper is organized into seven sections. Section II gives background on AL and surveys AL literature in remote sensing. Section III describes the adopted techniques for the implementation of uncertainty and diversity criteria and introduces the proposed sample labeling cost criterion. Section IV introduces the proposed optimization algorithms defined for the joint evaluation of the diversity and cost criteria. Section V illustrates the considered data set and the design of experiments, whereas Section VI shows the experimental results. Finally, Section VII draws the conclusion of this work.

II. ACTIVE LEARNING FOR REMOTE SENSING IMAGE CLASSIFICATION

In this section, we recall the general definition of AL, and then briefly survey AL techniques presented in the remote sensing literature. Let T be an initial training set with few labeled samples and U be a pool of unlabeled samples. AL techniques iteratively enrich the initial training set T by selecting the subset of most informative unlabeled samples from U for the considered classifier. Each iteration of AL consists of 3 steps: i) a batch X of informative unlabeled samples is selected by a query function, ii) then these samples are labeled by a supervisor (who assigns the true class

labels to the selected samples) and added to the current training set *T* and iii) finally the supervised classifier is retrained with the expanded training set. This procedure is iterated until a stop criterion is fulf illed. When the AL process is completed, the training set *T* includes a minimum number of highly informative samples (i.e., optimized training set) for the considered classifier. To end, the classifier is trained once more with the optimized training set, and then the image under investigation is classified [8].

The core of any AL method is related to the adopted sample selection strategy (i.e., query function). The basic approach to the selection of the most informative samples is to exploit an uncertainty criterion. As mentioned before, the uncertainty criterion aims at selecting the unlabeled samples that have the maximal uncertainty (i.e., the lowest confidence) on their correct class label among all unlabeled samples. The most uncertain samples are the most beneficial to be included in the training set due to the fact that they have the lowest probability to be accurately classified by the considered classifier. Uncertainty of samples can be defined in different ways depending on the considered classifier. In the remote sensing literature, several AL methods with different uncertainty measures have been presented. For example, in [4] the unlabeled sample that is closest to the classification boundary of each binary SVM in the one-against-all (OAA) multiclass architecture is considered as the most informative, and thus included in the current training set at each iteration of the AL process. An AL technique that selects the unlabeled sample maximizing the information gain is presented in [5]. In this work, the information gain is measured by the Kullback–Leibler divergence that is estimated between the posterior probability distribution of the current training set and the training set obtained in case of inserting each unlabeled sample, one by one, into the training set. In [6], the Entropy Query by Bagging (EQB) technique is proposed, which assesses the uncertainty of samples according to the maximum disagreement between a committee of classifiers. The disagreement among classifiers is measured by the entropy of the distribution of the different labels (obtained by a committee of classifiers). The samples with

maximum entropy are assigned as the most uncertain samples. A cluster assumption based AL method is presented in [7], which can overcome the problems related to the availability of biased initial training sets. This method exploits a histogram-thresholding algorithm to find out the most uncertain region in the one-dimensional SVM output space.

It is important to note that the use of only an uncertainty criterion is effective to select one sample at each iteration of AL, whereas it may result in poor performances in the case of choosing a batch of samples due to the possible redundancy (high similarity) between the selected samples. Thus, in order to select a batch of samples the query function should assess also the diversity of samples in addition to the uncertainty measure. To this end, AL methods that take into account the two criteria (uncertainty and diversity) have been recently presented in the remote sensing literature. For example, margin sampling by closest support vector method is presented in [6]. This method considers the smallest distance of each unlabeled sample to the hyperplanes associated to the binary SVMs in a OAA multiclass architecture as the uncertainty value of this sample. Then, at each iteration, the most uncertain unlabeled samples, which do not share the closest support vector, are added to the training set. In [8], different batch-mode AL techniques based on both uncertainty and diversity criteria for the classification of remote images with SVMs are presented. As an example, the Multiclass-Level Uncertainty with Enhanced Clustering Based Diversity (MCLU-ECBD) technique is proposed which initially selects the most informative unlabeled samples by the MCLU strategy, and then assesses the diversity of the most uncertain samples by a kernel-clustering technique. The *k*-means clustering in the kernel space is applied to the uncertain samples, and then the most uncertain sample of each cluster is included in the training set at each iteration of AL. The method presented in [7] is improved in [9] by including a diversity criterion to reduce the redundancy between the selected uncertain samples. Here, the diversity of uncertain samples is measured by using the same approach as [8].

All the above-mentioned methods may achieve high classification accuracy, thanks to the

optimized training set with a small number of highly informative labeled samples. However, they do not assess the cost of labeling of samples during the AL process. This may result in the selection of samples that are geographically far to each other or difficultly reachable by the supervisor, thus involving a high label acquisition cost. According to our knowledge, in the literature only few methods that consider the label acquisition costs during the AL process have been proposed [10], [11]. These methods aim to reduce the cost measured by the distance traveled during the labeling process. Accordingly, two variations of the methods are presented. The first one selects the most uncertain samples and defines the shortest path to travel among these samples according to the traveling salesman problem. The second one selects the samples that are closest to each other among the most uncertain samples by exploiting traveling salesman problem with profits (TSPP) [21]. Nonetheless, these methods have some important drawbacks: i) a diversity criterion is not considered in the selection process thus resulting in the possible selection of redundant samples; ii) the cost is modeled in a non-realistic way by ignoring the accessibility of samples on the ground; and iii)the distance between samples is calculated by a naive approach that only uses the two dimensional Cartesian coordinates neglecting both the altitude information and the possibility to use different transportation modes (i.e., foot or car) for moving from one sample to another. Thus, these methods do not model adequately the real applications constraints.

III. PROPOSED METHOD: CRITERIA FOR ACTIVE LEARNING

We propose a novel CSAL method that aims to select a batch $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_h\}$ of h samples at each iteration that are i) uncertain, ii) as more diverse as possible to each other, and iii) cost efficient. The proposed method is defined in the context of SVM classification problems by considering an one-against-all (OAA) architecture of binary SVMs for addressing multiclass problems. The proposed CSAL method assesses the uncertainty and diversity by using the MCLU [8] and angle based diversity (ABD) [12] techniques, respectively, which have been previously presented in the literature. Then, it introduces a cost criterion which is modeled by considering the

accessibility of each sample and the traveling time between successive samples. In the next subsections we briefly recall the MCLU and ABD techniques and then introduce the proposed cost criterion.

A. Uncertainty Criterion: Multiclass Level Uncertainty Technique

The uncertainty of samples is assessed by using the MCLU technique. This technique evaluates the confidence value $c(\mathbf{x})$ of each unlabeled sample $\mathbf{x} \in U$ with respect to the OAA SVM architecture, in which each binary SVM solves a problem defined by one information class against all the others [13], [14]. The confidence value $c(\mathbf{x})$ of each unlabeled sample is calculated according to its functional distance $f_i(\mathbf{x})$, $i = 1,...,r$ to the r decision boundaries of the binary SVM classifiers included in the OAA architecture. The confidence value of each unlabeled sample can be calculated using different strategies. Here, we use the difference function $c_{diff}(\mathbf{x})$ strategy that proved to be very effective in [8]. The $c_{diff}(\mathbf{x})$ strategy calculates the uncertainty between the two most likely classes by considering the difference between the first largest and the second largest distance values to the SVM hyperplanes [8]:

$$\begin{aligned} r_{1\max} &= \arg\max_{i=1,2,...,r} \{ f_i(\mathbf{x}) \} \\ r_{2\max} &= \arg\max_{j=1,2,...,r,\ j \neq r_{1\max}} \{ f_j(\mathbf{x}) \} \\ c_{diff}(\mathbf{x}) &= f_{r_{1\max}}(\mathbf{x}) - f_{r_{2\max}}(\mathbf{x}) \end{aligned} \tag{1}$$

If the $c_{diff}(\mathbf{x})$ value is small, the sample $\mathbf{x}$ is very close to the decision boundary between class $r_{1\max}$ and class $r_{2\max}$. In this case, since the decision for this sample is not reliable, it is considered as an uncertain pattern. On the opposite, if the $c_{diff}(\mathbf{x})$ value is high, the sample $\mathbf{x}$ is assigned to $r_{1\max}$ with high confidence and thus is not considered important for the AL procedure [8].

B. Diversity Criterion: Angle Based Diversity Technique

The proposed CSAL method assesses the similarity of the uncertain samples by the ABD technique. This technique measures the diversity of samples on the basis of the cosine angle distance, which is a similarity measure between two samples defined in the kernel space [12]:

$$\left|\cos\left(\angle(\phi(\mathbf{x}_i),\phi(\mathbf{x}_j))\right)\right| = \frac{\left|\phi(\mathbf{x}_i)\cdot\phi(\mathbf{x}_j)\right|}{\left\|\phi(\mathbf{x}_i)\right\|\left\|\phi(\mathbf{x}_j)\right\|} = \frac{K(\mathbf{x}_i,\mathbf{x}_j)}{\sqrt{K(\mathbf{x}_i,\mathbf{x}_i)K(\mathbf{x}_j,\mathbf{x}_j)}}$$
$$\angle(\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)) = \cos^{-1}\left(\frac{K(\mathbf{x}_i,\mathbf{x}_j)}{\sqrt{K(\mathbf{x}_i,\mathbf{x}_i)K(\mathbf{x}_j,\mathbf{x}_j)}}\right) \quad (2)$$

where $\phi(\cdot)$ is a nonlinear mapping function from the original feature space to a higher dimensional space and $K(\cdot,\cdot)$ is the kernel function that implicitly solves the dot product into the unknown transformed high dimensional space. The cosine angle distance in the kernel space can be calculated using only the kernel function $K(\cdot,\cdot)$ without direct knowledge of the mapping function $\phi(\cdot)$ [12]. The angle between two samples is small (cosine of angle is high) if these samples are close to each other and vice versa.

C. Proposed Sample Labeling Cost Criterion

The proposed CSAL method models the labeling cost of each sample by considering both its ground accessibility and the traveling time to visit the sample. Thus, the samples that are easily accessible (e.g., close to an infrastructure like a road) and require the shortest traveling time (i.e., are close to the previously visited samples) are considered as cost efficient (i.e., cheap) samples. To assess the accessibility of the samples and the cost of different trajectories associated with the selection of a batch of samples, we exploit the road network map and the DEM of the considered area. Moreover, the requirements with respect to the use of different transportation modes (e.g., foot or car) are also modeled.

It is worth noting that we model the cost of labeling batch of samples instead of that of labeling only one sample, since we assume that the supervisor visits a batch *X* of samples at each

iteration of the proposed CSAL method. The cost of sample labeling can be defined in terms of distance to be traveled by the supervisor or time taken by the supervisor to visit the samples or resources required by the transportation modes used by the supervisor (e.g., gas for car). In our work, we express the cost in terms of time. Accordingly, the total labeling cost t_l, i.e., the time, for the batch *X* of *h* samples selected at the *l*-th iteration is defined as:

$$t_l(X) = t_l^{initial} + t_l^{travel}(X) + t_l^{labeling} h \quad (3)$$

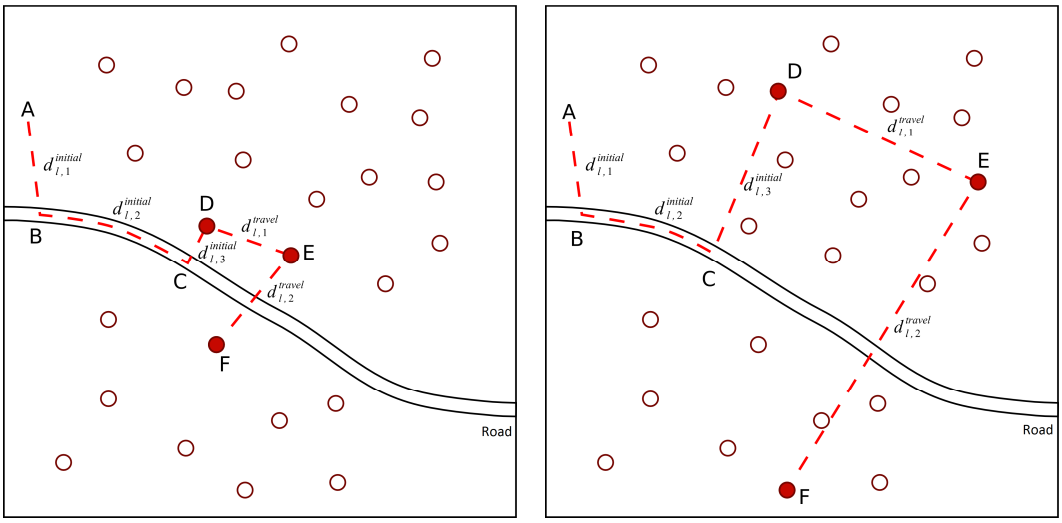
where $t_l^{initial}$ is the initial time to reach to the location of the first sample being labeled at the *l*-th iteration. $t_l^{initial}$ is estimated as $t_l^{initial} = d_l^{initial} / v$, where $d_l^{initial}$ is the traveling distance to the first sample being visited and v is the velocity of the considered transportation mode. The estimation of $t_l^{initial}$ depends on the transportation mode used for moving from the final location of the supervisor at the (*l*-1)-th iteration to that of the first sample being labeled. If the traveling time by car is shorter than that by foot, the supervisor travels by car and vice versa. In case of traveling by foot, $t_l^{initial}$ is estimated as $t_l^{initial} = (d_l^{initial} / v_{foot})$, and it is not necessary to use the road network map. However, in case of traveling by car, $d_l^{initial}$ is defined based on 3 distances: i) the distance $d_{l,1}^{initial}$ between the supervisor (i.e., his current location) and the road point where the car is left (this is done by foot), ii) the distance $d_{l,2}^{initial}$ from the initial road point to the final road point closest to the sample to be labeled (this is done by car), and iii) the distance $d_{l,3}^{initial}$ from the final road point to the sample being labeled (this is done by foot). Accordingly, $t_l^{initial}$ is estimated as $t_l^{initial} = (d_{l,1}^{initial} / v_{foot}) + (d_{l,2}^{initial} / v_{car}) + (d_{l,3}^{initial} / v_{foot})$, where the total traveling distance to reach the first sample being labeled is $d_l^{initial} = d_{l,1}^{initial} + d_{l,2}^{initial} + d_{l,3}^{initial}$.

The traveling time $t_l^{travel}(X)$ required to visit all the samples in *X* at the *l*-th iteration is estimated as $t_l^{travel}(X) = d_l^{travel}(X) / v$. $d_l^{travel}(X)$ is the shortest distance to travel between the *h*

samples and consists of the sum of the distances between each pair of samples. The shortest path to travel between the selected samples is estimated using the optimization algorithm presented in [10]. Here, we assume that traveling between the batch X of samples can be achieved by foot, i.e., $t_l^{travel}(X) = d_l^{travel}(X)/v_{foot}$, where v_{foot} is the velocity of traveling by foot. This is a reasonable assumption since selected samples can be most likely close to each other. However, it is straightforward to reformulate the cost function by considering a different assumption. The labeling time $t_l^{labeling}$ is the time taken by the supervisor to assign a label to each sample at the l-th iteration, and thus $t_l^{labeling}h$ is the total time for assigning labels to the h samples.

The distance between two samples is always estimated using also the altitude information of the samples provided by the DEM of the considered site in the Cartesian coordinates of the samples (i.e., geographic distance). In addition, the distances to the closest road points are obtained by exploiting the road network map of the considered area. Thanks to the road network map and the DEM data the proposed method for the definition of the labeling cost, unlike techniques proposed in [10]-[11], fits well with the real applications constraints.

According to the proposed sample labeling cost definition, the samples that have the shortest traveling time (those that are closest to each other) and require the shortest time to reach the road points (those that are easily accessible) are considered as cost efficient samples. In order to better understand this concept, Fig. 1 shows a qualitative example. Note that, for simplicity, the example is presented to visualize only the traveling routes for two different scenarios, which are related to low labeling cost [Fig. 1.a] and high labeling cost [Fig. 1.b]) in the case of selection of three samples. Thus, the uncertainty and diversity criteria are not considered in this figure.



(a) (b)

Fig. 1. The traveling routes defined by using the proposed sample labeling cost technique for two different scenarios related to: (a) low labeling cost, and (b) high labeling cost. A=initial location of the supervisor, F=final location of the supervisor. The locations being visited are given in the alphabetic order, and $d_l^{initial} = d_{l,1}^{initial} + d_{l,2}^{initial} + d_{l,3}^{initial}$ and $d_l^{travel} = d_{l,1}^{travel} + d_{l,2}^{travel}$.

IV. PROPOSED METHOD: OPTIMIZATION ALGORITHMS

At each iteration, the proposed CSAL method is based on the evaluation of the uncertainty, diversity and cost criteria applied in two consecutive steps to select the batch X of samples. In the first step, the $m > h$ most uncertain samples are selected according to the standard MCLU technique from a set U of unlabeled samples, whereas in the second step the most diverse and cheapest h samples among these m uncertain samples are chosen ($m > h > 1$).It is worth noting that the three criteria can be also jointly optimized. However, it will significantly increase the complexity of the estimations for the considered parameters. In order not to increase the complexity, we propose to use two consecutive steps. Accordingly, we define a criterion function made up of two terms: 1) a term that measures the diversity of the samples in the batch X, and 2) a novel term that evaluates the labeling cost of the batch X. In order to optimize the two terms, we propose two different algorithms that aim to find different tradeoffs between quality of the solution and computational time. The first algorithm achieves the optimization on the basis of a sequential forward selection (SFS) strategy, whereas the second one relies on a genetic algorithm (GA). Fig. 2 shows the block scheme of the proposed CSAL method.

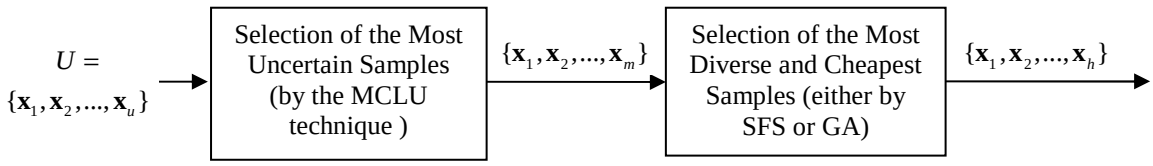


Fig. 2. Block diagram of the proposed cost-sensitive active learning (CSAL) method.

A. CSAL Method Optimized by a Sequential Forward Selection Strategy (CSAL-SFS)

The first algorithm performs the optimization by using a simple SFS strategy (CSAL-SFS). After selecting the *m* most uncertain samples according to the MCLU method, the uncertain samples that optimize the diversity and cost criteria based on the SFS strategy are chosen. According to the SFS strategy, the batch *X* is initially empty, and the *h* samples that are cheap and diverse to each other are sequentially chosen among the *m* uncertain samples. In this algorithm, the diversity and cost criteria are combined by using a weighting parameter λ. On the basis of this combination, a new sample $\mathbf{x}_t$ is included in the batch *X* at the *l*-th iteration according to the following optimization:

$$\mathbf{x}_t = \arg\min_{i=1,...,m} \left\{ \lambda t_l(X) + (1-\lambda) \left[\max_{x_j \in X} \frac{K(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{K(\mathbf{x}_i, \mathbf{x}_i) K(\mathbf{x}_j, \mathbf{x}_j)}} \right] \right\} \tag{4}$$

where λ provides the tradeoff between cost and diversity. The cosine angle distance between each uncertain sample $\mathbf{x}_i$, $i = 1, 2, ...m$ selected in the uncertainty step and the samples included in *X* are calculated and the maximum value is taken as the diversity value of the sample $\mathbf{x}_i$ (see the second part of (4)). The cost $t_l(X)$, which is obtained in case of including the sample $\mathbf{x}_i$ in *X* (i.e., $X = X \cup \mathbf{x}_i$), is calculated by (3). Then, the sum of the cost and diversity values weighted by λ is considered to obtain the combined value for $\mathbf{x}_i$. This value is calculated for each unlabeled sample selected in the uncertainty step. Then, the unlabeled sample $\mathbf{x}_t$ that minimizes such a value is included in *X*. This process is repeated until the number of samples of the set *X (*i.e., $|X|$*)* is equal

to *h*. If the λ value is large, the priority is given to the selection of "cheap" samples, whereas if it is small, the priority is given to the selection of diverse samples.

It is worth noting that the first selected sample in *X* directly affects the successive samples being selected, i.e., different solutions to *X* can be obtained depending on the initial sample. To select the best solution with respect to the initialization, we propose to use an exhaustive search, which aims to select each uncertain sample $\mathbf{x}_i$, $i = 1, 2, ...m$ sequentially as a first sample, and then to assess the quality of the *m* different solutions to *X*, i.e., $\{X_1, X_2, ..., X_m\}$. To this end, a criterion function *J* is defined. The criterion function value of the *k*-th solution $J_l(X_k)$ is computed as

$$J_l(X_k) = \lambda t_l(X_k) + (1-\lambda) D_l(X_k) \tag{5}$$

where $D_l(X_k)$ is the diversity value of the batch X_k that is calculated by exploiting the average of angle based distances computed in the kernel space between each pair of samples in X_k (see (2)) as

$$D_l(X) = \frac{2}{h \times (h-1)} \sum_{i=1}^{h-1} \sum_{j=i+1}^{h} \frac{K(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{K(\mathbf{x}_i, \mathbf{x}_i) K(\mathbf{x}_j, \mathbf{x}_j)}} \tag{6}$$

Then, the final batch *X* is selected as that which minimizes (5), i.e.,

$$\min_{k=1,2,...,m} \{J_l(X_k)\} \tag{7}$$

The labeling of the batch *X* of samples can be done by establishing a link between the classification system and the human expert on the ground. To this end, a connection to a remote server that provides geographical coordinates of the samples selected by the AL algorithm and a GPS should be available during the ground survey (or in alternative the AL algorithm can be run on a Laptop available on site during the ground data collection). After manual labeling, the selected samples are added to the current set.

This procedure is iterated until the desired number of samples is labeled, i.e., the upper bound of the cost for labeling samples is achieved. When the AL process is completed, the image

is classified by the considered SVM classifier. This is done by training the classifier with the final training set obtained at the end of the AL process. A general AL iteration of this method to select the batch X of h samples is summarized in Algorithm 1. It is worth noting that even if we use an exhaustive search in the initialization step, the SFS strategy cannot guarantee to find the optimal solution at each iteration as it does not consider any backtracking option.

Algorithm 1: CSAL-SFS

Inputs:

λ (weighting parameter that tunes the tradeoff between diversity and cost)
m (number of samples selected on the basis of their uncertainty)
h (batch size)

Output:

X (set of unlabeled samples to be included in the training set)

1. Initialize k=1.
2. Compute $c(\mathbf{x})$ for each sample $\mathbf{x} \in U$.
3. Select the set of m unlabeled samples with the lowest $c(\mathbf{x})$ value (most uncertain) $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m\}$.

Repeat

4. Initialize X_k to the empty set.
5. Include in X_k the k-th most uncertain sample.

Repeat

6. Compute the combination of diversity and cost with (4)
7. Include the unlabeled sample $\mathbf{x}_t$ which minimizes (4) in X.

Until $|X_k| = h$

8. k=k+1

Until $k = m$

9. Select the batch X that minimizes (5) with (7).
10. Add the labels to the set of samples $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_h\} \in X$ and include them in the current training set T.

B. CSAL Method Optimized by a Genetic Algorithm (CSAL-GA)

The second algorithm aims to improve the quality of the solution by optimizing the proposed CSAL method with a GA (CSAL-GA). GAs start with a predefined number of randomly generated initial solutions (i.e., chromosomes) and evolve the initial set of solutions (i.e., initial population) by: i) mutation; ii) cross-over; and iii) selection operations [16]-[18]. The crossover operation aims to generate new solutions by defining strategies to merge already available solutions, whereas the

mutation operator creates a new solution from each single solution independently from the other available solutions. After producing new solutions either by mutation or by cross-over operations, the best solutions are selected in the selection step, whereas the worst solutions are removed from the population. A fitness function is adopted to assess the quality of solutions. The process is repeated until a stopping criterion is satisfied. The reader is referred to [16]-[18] for detailed information on the GA theory.

In the CSAL-GA, after selecting the set of *m* most uncertain samples by the MCLU technique, the GA is adopted to select the optimum batch *X* of *h* samples that optimizes the aggregation of the diversity and labeling cost criteria as in (5). This algorithm includes 3 main steps: i) initialization of *n* solutions $\{X_1, X_2, ..., X_n\}$ by *n* times randomly selecting *h* samples among *m* samples; ii) generation of new solutions using either crossover or mutation operations, resulting in 2*n* solutions; and iii) elimination of the worst *n* solutions by selecting the best *n* solutions.

Let $\{X_1, X_2, ..., X_n\}$ be an initial set of solutions (i.e., initial population), where the *k*-th solution X_k is made up of *h* samples selected randomly among the *m* most uncertain samples. The proposed method initially creates a set of *n* new solutions by either crossover or mutation operations. In crossover operation, two solutions are randomly selected from the current population; then a new solution is generated by selecting samples from these solutions and by merging them. Let $X_k = \{\mathbf{x}_1^k, \mathbf{x}_2^k, ..., \mathbf{x}_h^k\}$ and $X_p = \{\mathbf{x}_1^p, \mathbf{x}_2^p, ..., \mathbf{x}_h^p\}$ be two randomly selected solutions (i.e., chromosomes) from the population. The new solution X_l is generated by selecting *r* samples from X_k and (*h*-*r*) samples from X_p, i.e., $X_l = \left\{\mathbf{x}_1^k, \mathbf{x}_2^k, ..., \mathbf{x}_r^k, \mathbf{x}_1^p, \mathbf{x}_2^p, ..., \mathbf{x}_{(h-r)}^p\right\}$. The samples from X_p are chosen randomly, whereas the *r* cheap and diverse to each other samples are selected among the *h* in X_k. To this end, the criterion function values for the all possible different

combinations of *r* samples are estimated by (5), and the batch of *r* samples which has the minimum criterion function value is selected.

In the mutation operation, a new solution is created from each solution X_k, $k = 1,2,...,n$. This is done by removing the weakest sample from X_k and inserting another sample (instead of the weakest sample), which is randomly selected from the set of *m* most uncertain patterns that were not included in X_k beforehand. In order to find the weakest sample, each sample $\mathbf{x}_i \in X_k$, $i = 1,2,...,h$, is sequentially eliminated from X_k, and then the criterion function value is estimated by (5). The sample for which the criterion function value is minimum is considered as the weakest sample.

After generating *n* new solutions by either mutation or cross-over operations, the criterion function value of each solution is calculated; then in the selection step the *n* best solutions that minimizes (5) are kept, whereas those having the highest criterion function values are discarded. The operations of mutation, crossover and selection are iterated until a stopping criterion is fulfilled, which is achieved when a solution *X* is found as the best solution within a predefined number of iterations. Once the stopping criterion is satisfied, the samples in *X* are added to the training set after manual labeling. This procedure is iterated until the desired number of samples is labeled, i.e., the maximum cost for labeling samples is reached. When the AL process is completed, the image is classified by the considered SVM classifier. Algorithm 2 summarizes a single AL iteration for the CSAL-GA.

It is worth noting that the capabilities of GA to explore the space of solutions in general results in selecting samples with a better quality than those identified with the SFS algorithm.

Algorithm 2: CSAL-GA

Inputs:

λ (weighting parameter that tunes the tradeoff between diversity and cost)
m (number of samples selected on the basis of their uncertainty)
h (batch size)

Output:

X (set of unlabeled samples to be included in the training set)

1. Compute $c(\mathbf{x})$ for each sample $\mathbf{x} \in U$.
2. Select the set of *m* unlabeled samples with lower $c(\mathbf{x})$ value (most uncertain) $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m\}$.
3. Create *n* different solutions $\{X_1, X_2, ..., X_n\}$ (each solution includes *h* samples randomly selected from the set of *m* most uncertain samples).

Repeat

4. Perform the mutation or cross-over operations to create the *n* new solutions.
5. Compute the criterion function value of each solution with (5).
6. Select the *n* solutions which minimize (5).

Until a solution X is selected as one of the best solutions during a predefined number of iterations.

7. Add the labels to the set of samples $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_h\} \in X$ and include them in the current training set *T*.

V. DATA SET DESCRIPTION AND DESIGN OF EXPERIMENTS

A. Data Set Description

Experiments were conducted on two multispectral images. The first data set is an image acquired by the Quickbird multispectral sensor on the city of Trento (Italy) in October 2005 (see Fig. 3). The selected test site is a section of 2066×3100 pixels with a spatial resolution of 0.7 m, and thus the size of the considered area of 1446×2170 m. The map of main roads and the DEM of the considered area (see Fig. 3) are available to assess the sample labeling cost. In order to show the performance of the proposed method for a larger area, the experiments were also carried out by assuming that the pixel spatial resolution is 7 m on the ground when calculating the distance between the samples. This changes completely the relations between the different variables considered in the AL process and thus defines a problem which is significantly different form the first one. Accordingly, the simulated area is 14460×21700 m. Thus we have two scenarios. In the first scenario (Scenario 1) each pixel of the image is considered as associated to an area on the ground of 0.7 m, whereas in the second one (Scenario 2) each pixel of the image is considered as

associated to an area on the ground of 7 m. The available ground reference samples (5785 samples) were randomly divided to derive a validation set *V* of 195 samples (which is used for model selection), a test set *TS* of 2902 samples (which is used for accuracy assessment), and a pool of 2688 samples. The 2% of the samples of each class in the pool are randomly selected as initial training samples *T* and the rest are considered as unlabeled samples *U*. Table I shows the land cover classes and the related number of samples used in the experiments.

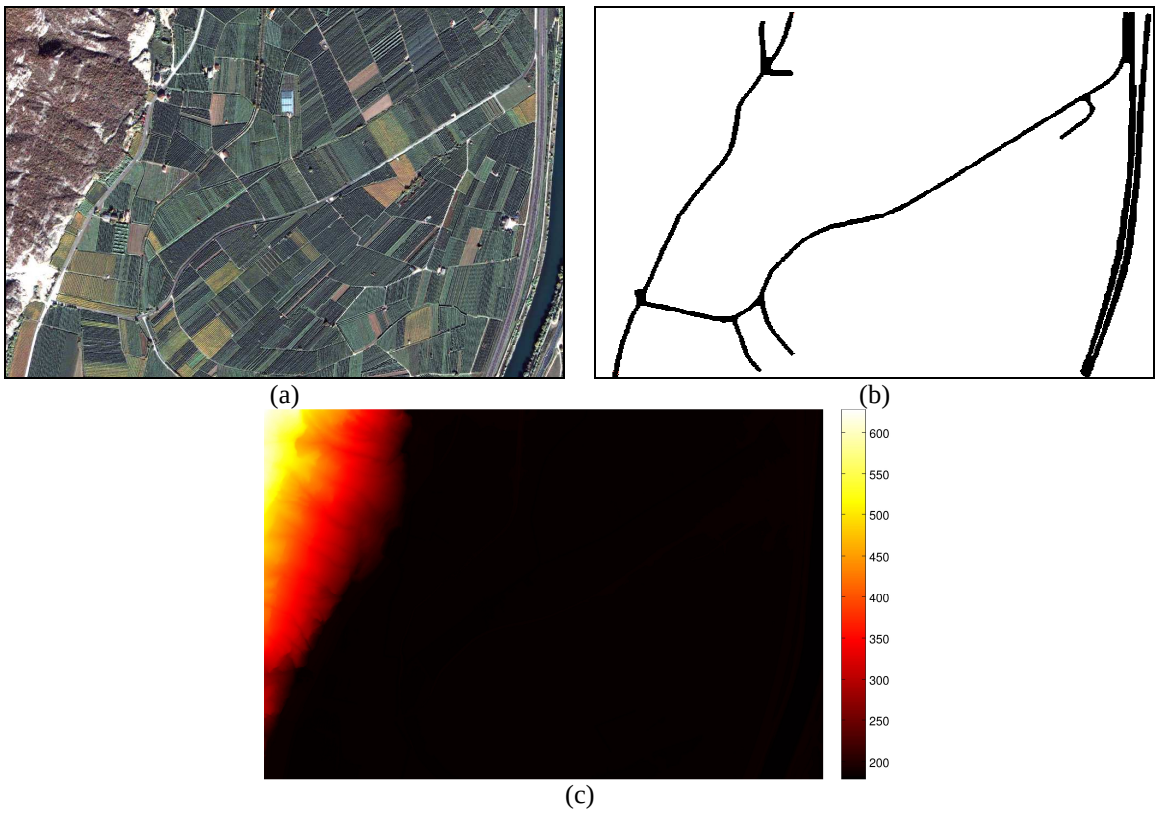


Fig. 3. Trento Quickbird data set: (a) true color composite, (b) map of the main roads and (c) DEM of the considered area.

TABLE I. NUMBER OF SAMPLES OF EACH CLASS IN THE INITIAL TRAINING SET (*T*), THE UNLABELED SAMPLE SET (*U*), THE VALIDATION SET (*V*), AND THE TEST SET (*TS*) FOR THE TRENTO QUICKBIRD DATA SET

Land-cover classes	***T***	***U***	***V***	***TS***
Water	8	383	28	531
Asphalt	12	565	42	602
Field	14	659	49	732
Forest	18	802	60	820
Bare soil	5	222	16	217
Total	57	2631	195	2902

The second data set is a multispectral image acquired by the GeoEye system on a larger area of the city of Trento (Italy) than in the previous case in September 2011 (see Fig. 4). The map of main roads and the DEM of the considered area (see Fig. 4) are available to assess the sample labeling cost. The available ground reference samples (16079 samples) were randomly divided to derive a validation set *V* with 231 samples, a test set *TS* with 8210 samples and a pool with 7638 samples. The 0.8% of the samples of each class in the pool are randomly selected as initial training samples for a total of 63 samples, and the rest are considered as unlabeled samples *U*. Table II shows the land cover classes and the related number of samples used in the experiments.

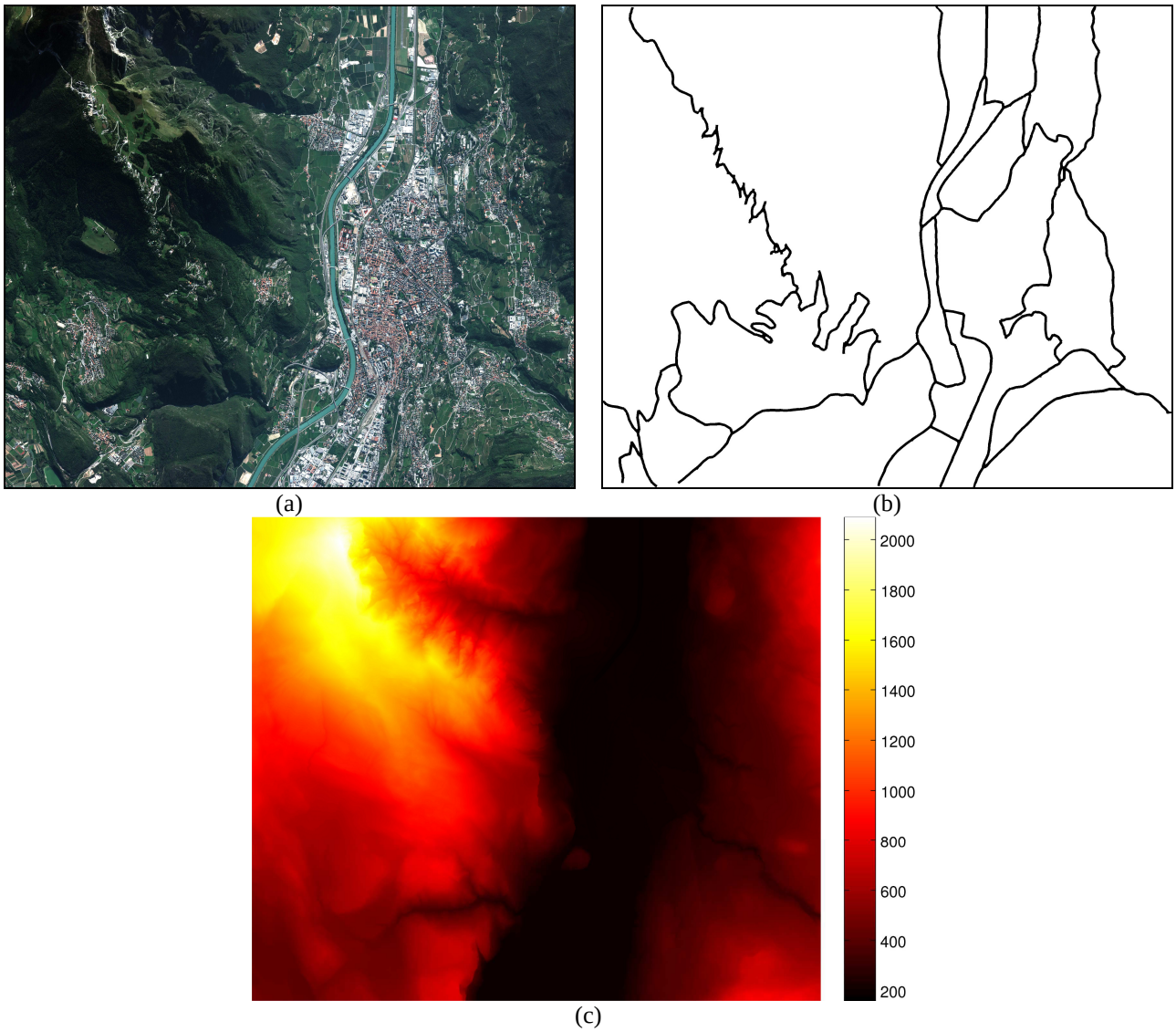


Fig. 4. Trento GeoEye data set: (a) true color composite, (b) map of the main roads and (c) DEM of the considered area.

TABLE II. NUMBER OF SAMPLES OF EACH CLASS IN THE INITIAL TRAINING SET (*T*), THE UNLABELED SAMPLE SET (*U*), THE VALIDATION SET (*V*), AND THE TEST SET (*TS*) FOR THE TRENTO GEOEYE DATA SET

Land-cover classes	***T***	***U***	***V***	***TS***
Trees	12	1525	45	1635
Fields	13	1595	47	1630
Asphalt	12	1488	44	1575
Water	4	490	14	609
Shadow	7	870	26	1035
Buildings	13	1616	48	1726
Total	62	7579	228	8210

B. Design of Experiments

In our experiments, we used an SVM classifier with Radial Basis Function (RBF) kernel [14]. The values for the regularization parameter *C* and the spread γ of the RBF kernel parameters were chosen performing a grid-search model selection only at the first iteration of the AL process as suggested in [8].

In our experiments, the velocity of traveling by foot was set to 6 km/hours, whereas that of traveling by car was fixed to 50 km/hours. In addition, $t_{labeling}$ was set to 2 minutes. The value of *m* (number of samples selected in the uncertainty step) was selected equal to 80, whereas the value of *h* (number of samples being selected at each iteration of AL) was chosen equal to 5. The value of λ was varied as $\lambda = 0.2,0.5,0.8$. All the distances are estimated by the Euclidean distance between the samples. The shortest distance d_{travel} required to travel between the selected samples is estimated using the optimization algorithm presented in [10], [20]. The distance $d_{l,2}^{initial}$ (from the initial road point to the final road point closest to the sample being labeled) is estimated using the method presented in [15]. The value of *n* for the GA (the number of initial solutions) is set equal the value of *m*. The stopping criterion for the CSAL-GA is achieved when the same solution is selected as one of the best solutions during the last 5 iterations.

We compared the proposed CSAL-SFS and CSAL-GA with the state-of-the-art AL techniques, i.e. the MCLU-ECBD [8] technique and the spatially cost-sensitive AL technique [10].

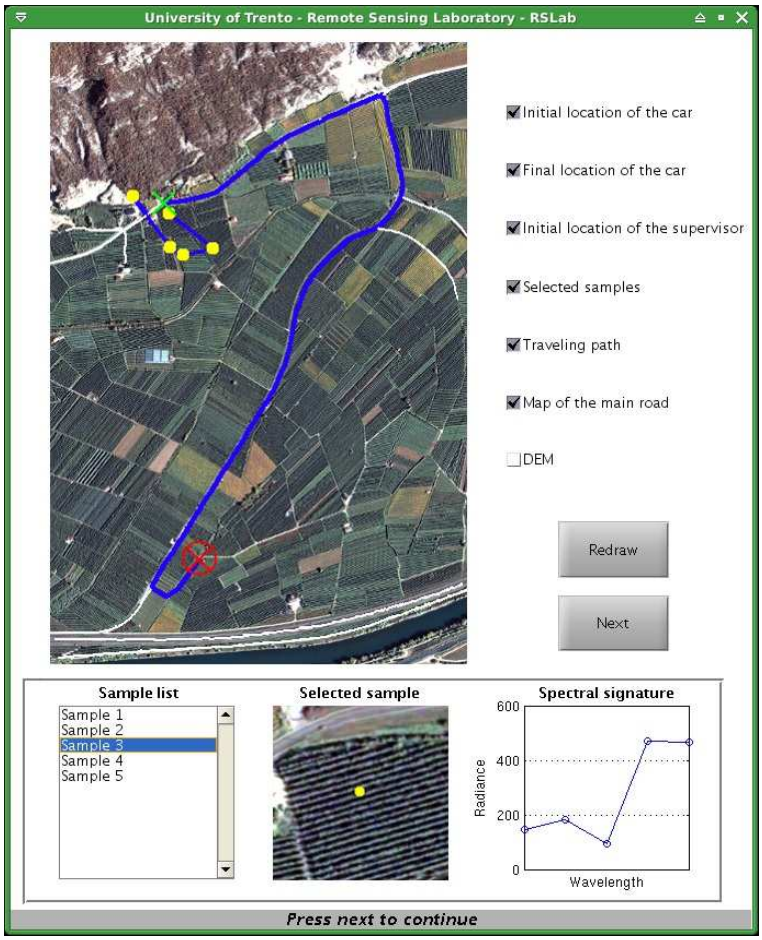


Fig. 5. Developed user interface, which shows the traveling route defined by using the proposed CSAL-GA method together with the selected five samples on the Quickbird data set. red circle=initial location of the supervisor; green cross=final location of the supervisor; yellow circles: selected samples; blue line: the path to be followed by the supervisor.

The MCLU-ECBD technique is implemented by initially selecting *m* most uncertain samples by the MCLU technique, and then choosing *h* diverse samples among the *m* samples by the ECBD technique. In order to implement the method in [10], the MCLU technique is used for the selection of the *m* uncertain samples, and then the TSPP is applied to these samples for the selection of the *h* cost sensitive samples as in [10]. This method is denoted as MCLU-TSPP in the experiments. It is worth emphasizing that the MCLU-ECBD technique considers the uncertainty and diversity criteria, whereas the MCLU-TSPP includes the uncertainty and cost criteria. We also compared the results with random sampling method that selects the samples randomly at each iteration without exploiting any AL criterion.

All experimental results are referred to the average accuracies obtained in ten trials according to ten initial randomly selected training sets. Results are reported as learning rate curves, which show the average classification accuracy versus i) the total time (i.e., cost of sample labeling) spent during the collection of ground reference data, and ii) the number of labeled samples. For a fair comparison, the total time was calculated using the proposed labeling cost method for all the methods used in the experiments.

The experiments were done by implementing the proposed technique in a Matlab software tool having the interface shown in Fig. 5. The figure shows an example of the traveling route defined by using the proposed CSAL-GA method and the selected samples on the Quickbird data set together with the initial and final locations of the supervisor for one AL iteration.

VI. EXPERIMENTAL RESULTS

We did different kinds of experiments aimed to: i) perform a sensitivity analysis with respect to λ values, and ii) compare the effectiveness of the proposed CSAL-SFS and CSAL-GA between each other and also with the MCLU-ECBD and the MCLU-TSPP techniques for both data sets.

A. Sensitivity analysis to different λ values

We analyzed the performances of the CSAL-SFS and the CSAL-GA versus the value of λ. As an example, Fig. 6 and Fig. 7 compare the overall accuracies versus the labeling time (i.e., the cost) obtained by the CSAL-SFS and the CSAL-GA for the Quickbird data set (both scenarios) and the GeoEye data set, respectively. From the figures, one can see that selecting higher values of λ results in better classification accuracies compared to those obtained by small values of λ for both scenarios. This is due to the fact that small λ values result in the selection of more diverse samples with the possible drawback of increasing time. This behavior is more evident in the case of the second scenario of the Quickbird and the GeoEye data sets for both the CSAL-SFS and the CSAL-GA (see Fig. 6.b, Fig. 6.c , Fig. 7.b and Fig. 7.c). As an example, the CSAL-GA with

$\lambda = 0.8$ provides an accuracy of 92.05% when 10 hours are spent for the collection of ground reference data, whereas the accuracies obtained by the same method in the same time when $\lambda = 0.2$ and $\lambda = 0.5$ are 89.06% and 90.03%, respectively (see Fig. 7.b). Thus, the value of the parameter λ defined by the user is crucial. Another interesting observation is that using high values of λ leads to a faster convergence than when using small values of λ. Note that the reason of achieving higher accuracies with the proposed AL method compared to the case of using the whole pool as training set is related to the presence of noisy samples (or outliers) in the pool. These samples do not properly model the distribution of test pixels. It is worth nothing that an outlier is expected to be assigned to a wrong class by the classifier with high confidence (i.e., with low uncertainty); accordingly, it is not selected as an uncertain sample by the proposed AL method.

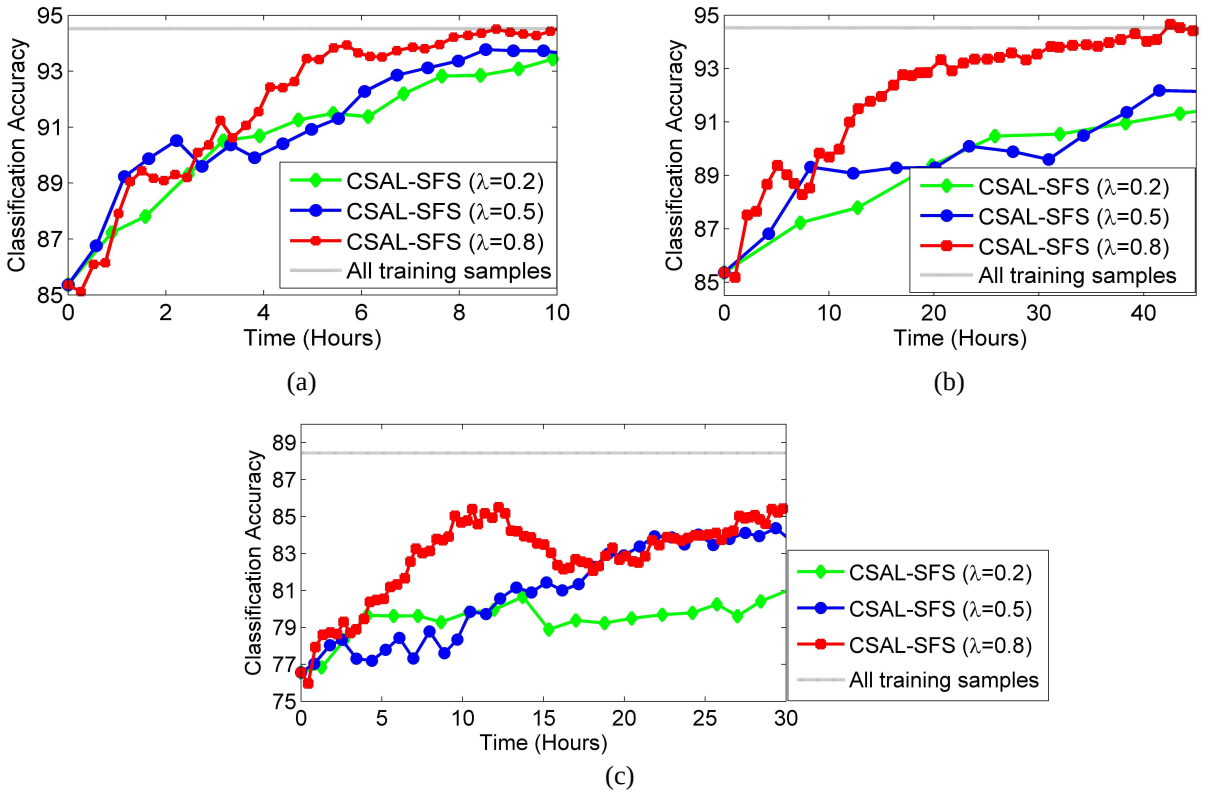


Fig. 6. Average (on ten trials) overall classification accuracy (in %) versus the total labeling time obtained by the proposed CSAL-SFS method for different values of λ for (a) the first scenario of the Quickbird data set, (b) the second scenario of the Quickbird data set, and (c) the GeoEye data set. The dashed line "All training samples" shows the accuracy obtained including all unlabeled samples in the training set after manual labeling.

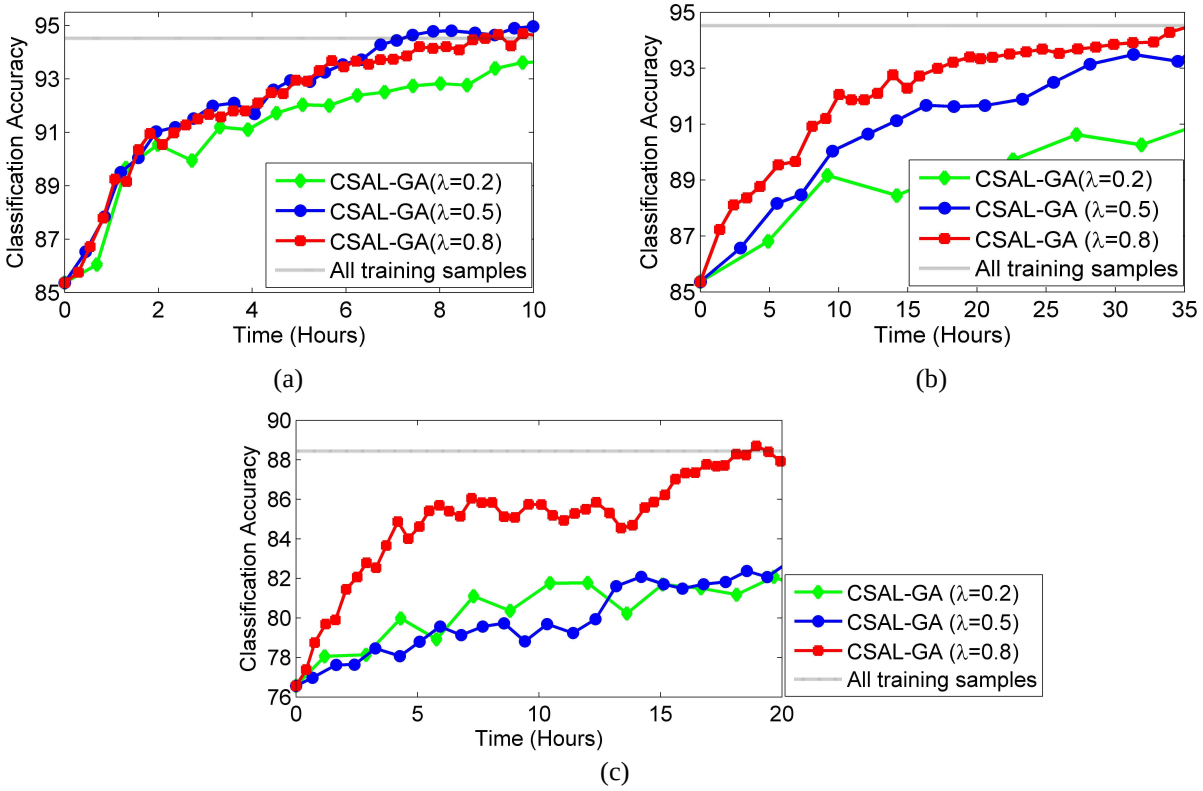


Fig. 7. Average (on ten trials) overall classification accuracy (in %) versus the total labeling time obtained by the proposed CSAL-GA method for different values of λ for (a) the first scenario of the Quickbird data set, (b) the second scenario of the Quickbird data set, and (c) the GeoEye data set. The dashed line "All training samples" shows the accuracy obtained including all unlabeled samples in the training set after manual labeling.

It is worth noting that the effect of λ is opposed in the case of comparing the classification accuracy versus only the number of labeled samples by neglecting the labeling time. As examples, for the Quickbird data set, Fig. 8 and Fig. 9 show the results obtained by the CSAL-SFS and CSAL-GA, respectively, varying the λ values. From the figures, one can observe that the CSAL-SFS and CSAL-GA select smaller number of samples to be labeled when small values of λ are considered for both scenarios. This is due to the fact that small λ values provide the selection of more diverse samples, which results in the need to label a smaller number of samples to reach convergence. As an example, the CSAL-SFS reaches an accuracy of 92.19% with 102 samples when $\lambda = 0.2$, whereas the accuracies obtained with the same number of samples when $\lambda = 0.5$ and $\lambda = 0.8$ are 90.92% and 89.30%, respectively (Fig. 8.b). In the case of CSAL-GA, the results obtained when $\lambda = 0.2$ and $\lambda = 0.5$ are very similar to each other, and outperform those obtained when $\lambda = 0.8$. For the Quickbird data set, Table III and Table IV report the labeling time (in

hours) and the number of labeled samples for achieving similar accuracies with respect to different λ values obtained by the CSAL-SFS and CSAL-GA, respectively. From the tables, one can observe that in the case of small values of λ the time taken by both the CSAL-SFS and the CSAL-GA is large (due to the priority in the selection of samples having high distance among them) while the number of labeled training samples is smaller. On the other hand, increasing the λ value results in a shorter time even if more samples are labeled (due to the priority in the selection of samples that are closer to each other). This analysis significantly shows that focusing only on the number of labeled samples for assessing the effectiveness of AL techniques can be misleading.

The inclusion of the cost of labeling term (expressed in time) points out that in operational application it can be more convenient (i.e., faster) to label a larger number of samples that are close each other than a smaller number of samples that are far each other. It is worth noting that the same behavior is also observed in the results obtained for the GeoEye data set (not reported for space constraints).

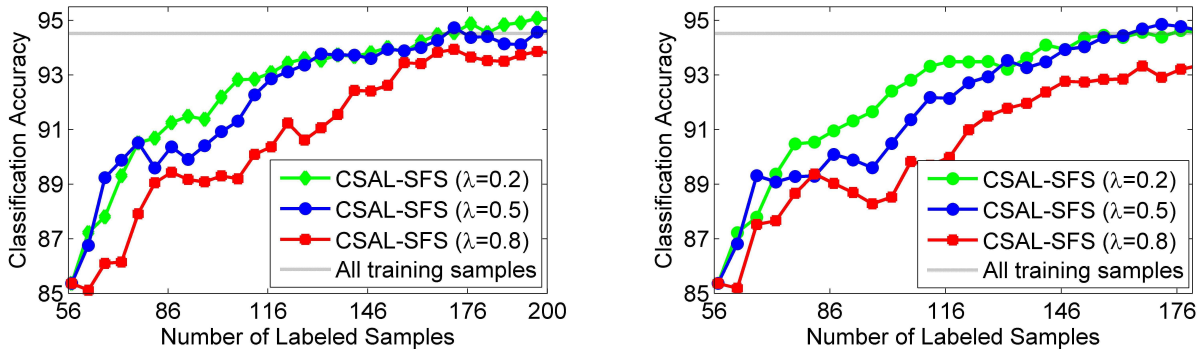
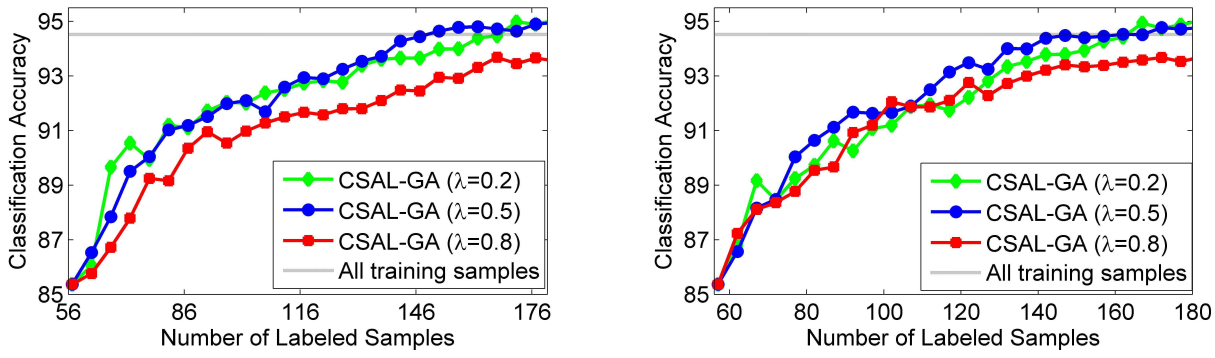


Fig. 8. Average (on ten trials) overall classification accuracy (in %) versus the number of labeled samples obtained by the proposed CSAL-SFS method for different values of λ for (a) the first scenario and (b) the second scenario of the Quickbird data set. The dashed line "All training samples" shows the accuracy obtained including all unlabeled samples in the training set after manual labeling.



(a) (b)

Fig. 9. Average (on ten trials) overall classification accuracy (in %) versus the number of labeled samples obtained by the proposed CSAL-GA method for different values of λ for (a) the first scenario and (b) the second scenario of the Quickbird data set. The dashed line "All training samples" shows the accuracy obtained including all unlabeled samples in the training set after manual labeling.

TABLE III. LABELING TIME (IN HOURS) AND NUMBER OF LABELED SAMPLES TAKEN FOR OBTAINING SIMILAR ACCURACIES BY THE CSAL-SFS TECHNIQUE WITH RESPECT TO DIFFERENT λ VALUES FOR THE FIRST SCENARIO OF THE QUICKBIRD DATA SET

λ	Scenario 1			Scenario 2		
	Time	Samples	Accuracy	Time	Samples	Accuracy
0.2	16	162	94.22	101	142	94.10
0.5	13	167	94.27	74	152	94.03
0.8	7	217	94.31	37	242	94.09

TABLE IV. LABELING TIME (IN HOURS) AND NUMBER OF LABELED SAMPLES TAKEN FOR OBTAINING SIMILAR ACCURACIES BY THE CSAL-GA TECHNIQUE WITH RESPECT TO DIFFERENT λ VALUES FOR THE SECOND SCENARIO OF THE QUICKBIRD DATA SET

λ	Scenario 1			Scenario 2		
	Time	Samples	Accuracy	Time	Samples	Accuracy
0.2	13	162	94.39	86	162	94.27
0.5	8	152	94.28	42	148	94.38
0.8	7	217	94.21	33	207	94.28

B. Comparison among the Proposed and the Literature Methods

In this sub-section we carried out two sets of experiments. On the basis of the analysis done in the previous subsection, here we compared the accuracies versus only the time (i.e., the cost). In the first set of experiments, we compared the effectiveness of the proposed CSAL-SFS and CSAL-GA between them. Fig. 10 shows the overall accuracies versus the time obtained by the CSAL-SFS and the CSAL-GA for both scenarios of the Quickbird data set and the GeoEye data set, respectively. In the figure, we reported the highest average accuracy obtained with the best values

of the parameters λ. For the first scenario of the Quickbird data set , the highest accuracies for CSAL-SFS are obtained with $\lambda = 0.8$, whereas those for CSAL-GA are obtained for λ=0.5. For the second scenario of this data set and the GeoEye data set, the highest accuracies for both CSAL-SFS and CSAL-GA are obtained when λ=0.8. For both data sets, one can observe that the CSAL-GA provides a more effective selection of samples than the CSAL-SFS. In other words, it achieves slightly higher accuracies with the same cost (or the same accuracy with less cost). On the contrary, the CSAL-SFS results in a slightly lower computational complexity than the CSAL-GA. In our experiments, for the Quickbird data set the computational time taken for a single AL iteration from the CSAL-SFS is 7.79 seconds, whereas that required by the CSAL-GA is 9.27 seconds. For the GeoEye data set, the computational time taken for a single AL iteration from the CSAL-SFS is 16.95 seconds, whereas that required by the CSAL-GA is 24.48 seconds (note that this time significantly depends on the choice of the GA parameters). Nonetheless, this time difference between the techniques is very small and both are suitable to be used on a real operational scenario where these computation times are negligible when compared with the sample labeling time (*i.e.*, cost) .

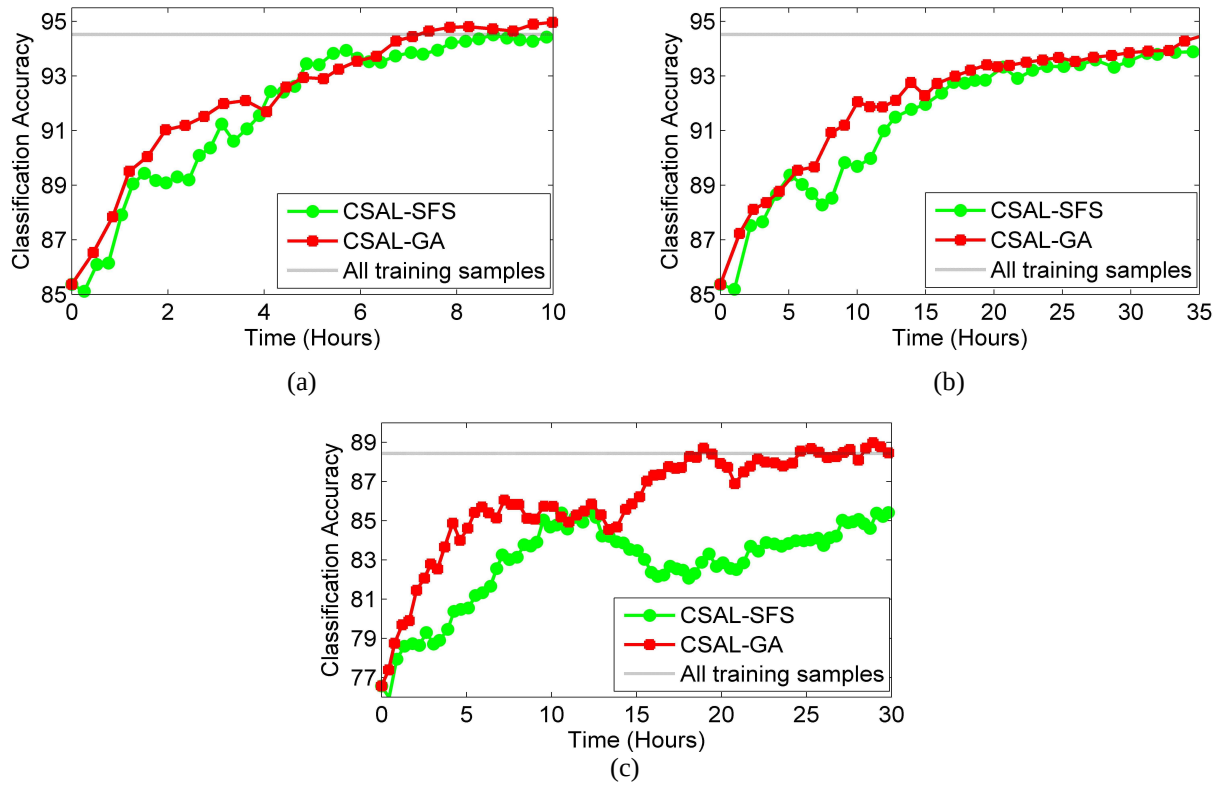
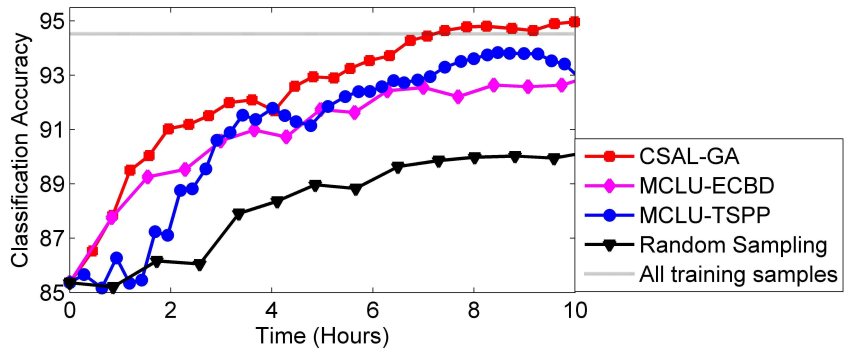


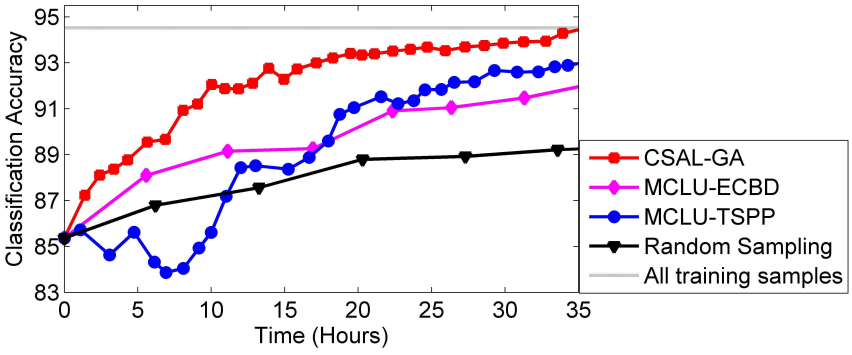
Fig. 10. Average (on ten trials) overall classification accuracy (in %) obtained by the proposed methods using (a) the first scenario of the Quickbird data set, (b) the second scenario of the Quickbird data set, and (c) the GeoEye data set. The dashed line "All training samples" shows the accuracy obtained including all unlabeled samples in the training set after manual labeling.

Finally, we compared the proposed CSAL-SFS and CSAL-GA with the state of the art AL techniques, i.e., the MCLU-ECBD [8], the MCLU-TSPP [10] and also with the random sampling. Fig. 11 shows the average overall accuracies versus the labeling time obtained for both scenarios of the Quickbird data set and the GeoEye data set, respectively. In the figure, we only report the results obtained by the CSAL-GA only, due to its higher classification accuracies compared to the CSAL-SFS. By analyzing the figure, one can observe that the CSAL-GA leads to the highest accuracies for all the iterations and significantly outperforms the MCLU-ECBD and the MCLU-TSPP methods for both data sets. As an example, on the first scenario of the Quickbird data set, the CSAL-GA provides an accuracy of 94.96% when 10 hours are spent for the collection of ground reference data, whereas those obtained by the MCLU-ECBD, the MCLU-TSPP and random sampling under the same time, are 92.63%, 92.98% and 90.21%, respectively (see Fig. 11.a). Moreover, the CSAL-GA provides an accuracy of 94.66% spending only 7 hours for the label collection process, whereas the MCLU-ECBD and the MCLU-TSPP methods require 20 and 14 hours to achieve a similar accuracy, respectively (see Fig. 11.a). The accuracy differences between the proposed the CSAL-GA and the other methods are higher in the case of second scenario of the Quickbird data set and the GeoEye data set (see Fig. 11.b and Fig. 11.c). For example, the CSAL-GA provides an accuracy of 92.05 % spending only 10 hours for the label collection process, whereas those obtained by the MCLU-ECBD, the MCLU-TSPP and random sampling in the case of the same are 89.02%, 85.60% and 87.01%, respectively (see Fig. 11.b). Another important result is that the CSAL-GA reaches convergence with smallest labeling time for both data sets. In greater details, for both data sets the proposed CSAL-GA is more effective than the MCLU-TSPP, which is one of the few cost based AL method presented so far in the remote sensing literature. The higher performance of the proposed CSAL-GA with respect to MCLU-TSPP relies on: i) considering the diversity criterion, and ii) modeling the cost in a more reliable way taking into account real application constraints (by considering the requirement to use

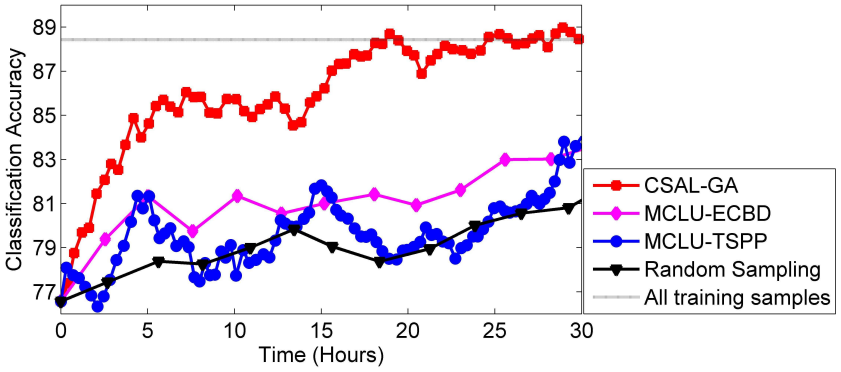
different transportation modes and by exploiting the road map and DEM data). Moreover, the proposed method significantly outperforms the MCLU-ECBD due to the fact that the latter does not consider any cost criterion for the selection of the samples. Note that also in this case the reason of achieving higher accuracies with the AL methods compared to the case of using the whole pool as training set is related to the presence of noisy samples (or outliers) in the pool.



(a)



(b)



(c)

Fig. 11. Average (on ten trials) overall classification accuracy (in %) obtained by the CSAL-GA, the MCLU-ECBD, the MCLU-TSPP and random sampling methods using (a) the first scenario of the Quickbird data set, (b) the second scenario of the Quickbird data set, and (c) the GeoEye data set. The dashed line "All training samples" shows the accuracy obtained including all unlabeled samples in the training set after manual labeling.

VII. DISCUSSION AND CONCLUSION

In this paper, we have presented a novel cost-sensitive AL (CSAL) method to the definition of effective training sets for the classification of remote sensing images with SVMs. The proposed CSAL method evaluates the uncertainty, diversity and cost criteria in order to select a batch of most informative samples to be labeled and included in the training set. The uncertainty of samples is assessed by the MCLU technique, whereas the diversity is measured based on cosine angle distances between the samples. In the proposed method, the sample labeling cost is expressed in terms of time and defined using a novel method. This method models the cost according to both samples ground accessibility (e.g., time required to reach to an infrastructure like a road) and traveling time between the samples. We take into account the use of different transportation modes with different properties (e.g., foot or car) for moving from a sample to the other in the computation of the cost. To this end, the road map and the DEM of the considered area are exploited. It is worth noting that, thanks to the DEM (that provides the altitude information in addition to the two dimensional Cartesian coordinates of the samples), the geographic distance is calculated between the samples. Moreover, thanks to the road map, the traveling distances are reliably estimated taking into account the precise position of the road.

In order to evaluate the above-mentioned three criteria (i.e., the uncertainty, diversity and cost), a two steps procedure is adopted in the paper. The first step is devoted to the selection of the most uncertain samples, whereas the second step aims at choosing the cheapest and most diverse samples among the uncertain ones. This can be done by using two different algorithms, which differ from each other with respect to the adopted optimization strategy. The first algorithm (i.e., the CSAL optimized by SFS) is simple and very fast and exploits the sub-optimal sequential forward selection strategy, whereas the second one (i.e., the CSAL optimized by GA) achieves the optimization on the basis of the genetic algorithm. In the experiments, we compared the performance of these optimization algorithms. By this analysis, we observed that, as expected, the

CSAL-GA yields better accuracies under the same labeling cost with respect to the sub-optimal CSAL-SFS. This is achieved by slightly increasing the computational time necessary to find the solution. Nonetheless, this time is negligible with respect to the time taken for collecting labels.

In the experimental analysis, we also compared the proposed method with the most promising state-of-the-art AL methods presented in the remote sensing literature. By this comparison, we observed that the proposed method allows one to significantly reduce the cost of the collection of reference samples to reach the desired classification accuracy compared to the state of the art AL methods. From another perspective, it can achieve the same accuracy reached by other techniques with a sharply smaller labeling cost. We underline that this is a very important advantage, because the main goal of AL is to optimize the training set with a minimum cost. An important issue pointed out from our result is that when label collection is done on the ground, minimizing the number of labeled samples should not be the goal of AL technique. Indeed, fixed a target accuracy, it can be more efficient to collect a large number of labeled samples that are close to each other (and only partially diverse) than a small number of labeled samples that are far to each other (and very diverse).

It is worth noting that proposed method is intrinsically classifier independent. Even if here it is implemented in the framework of the SVM classifier (because of its efficiency for remote sensing image classification), it can be easily adopted for the other classifiers. This can be done by selecting suitable techniques for assessing uncertainty and diversity of samples in the framework of the considered classifier, and then using them in the framework of the algorithms presented in this paper.

As a final remark, we would like to point out that the use of efficient techniques for the exploitation of AL methods in real applications is becoming a more and more important topic. In this context, the proposed method is very promising as it allows optimizing the definition of a training set, decreasing significantly the cost and effort required for reference data collection. As a

future development of this work, we plan to apply proposed technique and the related software tool for a real label collection task on the ground. Moreover, we plan to improve the initialization of the proposed method,, which can affect the convergence time of the process especially for large images.

REFERENCES

[1] L. Bruzzone, M. Chi, M. Marconcini, "A novel transductive SVM for the semisupervised classification of remote-sensing images", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 11, pp. 3363-3373, 2006.

[2] G. Schohn and D. Cohn, "Less is More: Active Learning with Support Vector Machines", *Proc. 17th Int'l Conf. Machine Learning (ICML '00)*, pp. 839-846, 2000.

[3] S. Tong and D. Koller, "Support Vector Machine Active Learning with Applications to Text Classification", *Proc. 17th Int'l Conf. Machine Learning (ICML '00)*, pp. 999-1006, 2000.

[4] 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.

[5] 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.

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

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

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

[9] S. Patra, L. Bruzzone, A cluster-assumption based batch mode active learning technique, *Pattern Recognition Letters*, vol. 33, 2012, pp. 1042-1048.

[10] A. Liu, G. Jun and J. Ghosh, "Spatially cost-sensitive active learning," *In SIAM International Conference on Data Mining*, Sparks, Nevada, USA, pp.814-825, 2009.

[11] A. Liu, G. Jun, and J. Ghosh, "Active learning of hyperspectral data with spatially dependent label acquisition costs," *IEEE International Geoscience and Remote Sensing Symposium*, Cape Town, South Africa, pp. V-256 - V-259, 2009.

[12] K. Brinker, "Incorporating Diversity in Active Learning with Support Vector Machines," *Proceedings of the International Conference on Machine Learning*, Washington DC, pp. 59-66, 2003.

[13] F. Melgani, L. Bruzzone, "Classification of Hyperspectral Remote Sensing Images With Support Vector Machines," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778- 1790, Aug. 2004.

[14] 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.

[15] E. W. Dijkstra, "A note on two problems in connexion with graphs," Numerische Mathematik, vol. 1, no. 1, pp. 269-271, 1959.

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

[17] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs. New York: Springer-Verlag, 1992.

[18] S. Bandyopadhyay and S. K. Pal, Classification and Learning Using Genetic Algorithms: Application in Bioinformatics and Web Intelligence. Berlin, Germany: Springer, 2007.

[19] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[20] D. S. Johnson and L. A. McGeoch. The traveling salesman problem: A case study. In E. H. Aarts and J. K. Lenstra, editors, Local search in combinatorial optimization, pages 215–310. Wiley, 1997.

[21] M. Gendreau, G. Laporte, and F. Semet. A branch-and-cut algorithm for the undirected selective traveling salesman problem. Networks, 32:263–273, 1998.