Title: Active Learning for Domain Adaptation in the Supervised Classification of Remote Sensing Images

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

Date of Publication: 2012

Author(s): Claudio Persello and Lorenzo Bruzzone

Volume: 50, Issue: 11

Page(s): 4468 - 4483

DOI: 10.1109/TGRS.2012.2192740

ACTIVE LEARNING FOR DOMAIN ADAPTATION IN THE SUPERVISED CLASSIFICATION OF REMOTE SENSING IMAGES

Claudio PERSELLO, *Member IEEE* and Lorenzo BRUZZONE, *Fellow IEEE*

Dept. of Information Engineering and Computer Science, University of Trento,

Via Sommarive, 14, I-38123 Trento, Italy;

e-mail: claudio.persello@disi.unitn.it, lorenzo.bruzzone@ing.unitn.it.

Abstract— This paper presents a novel technique for addressing domain adaptation (DA) problems with active learning (AL) in the classification of remote sensing images. DA models the important problem of adapting a supervised classifier trained on a given image (source domain) to the classification of another similar but not identical image (target domain) acquired on a different area. The main idea of the proposed approach is to iteratively labeling and adding to the training set the minimum number of the most informative samples from the target domain, while removing the source-domain samples that do not fit with the distributions of the classes in the target domain. In this way, the classification system exploits already available information, i.e., the labeled samples of source domain, in order to minimize the number of target domain samples to be labeled, thus reducing the cost associated to the definition of the training set for the classification of the target domain. In addition, we define a convergence criterion that allows the technique to stop the iterative AL process on the target domain without relying on the availability of a test set for it. This is an important contribution, as in operational applications it is not realistic to assume that a test set for the target domain is available. Experimental results obtained in the classification of very high resolution (VHR) and hyperspectral images confirm the effectiveness of the proposed technique.

Index Terms— Active learning, automatic classification, domain adaptation, hyperspectral data, remote sensing, transfer learning.

I. INTRODUCTION

In the last years, the advances in remote-sensing technology have led to a growing availability of space-borne data giving the opportunity to develop several important applications related to land-cover monitoring and mapping. This great opportunity poses the problem to develop adequate classification systems capable to produce accurate land-cover maps at reasonable cost and time. At the present, the most common automatic classification techniques used for obtaining land-cover maps from remotely sensed images are based on supervised learning methods, which rely on a set of labeled reference samples for training the classification algorithm. However, these methods require a set of new training samples every time that a new remote sensing image has to be classified, leading to high costs and critical constraints for the acquisition of new reference information. Hence, an important aspect to be considered is the development of tools for defining adequate and reliable training sets. Such tools may work synergically with the user and the classification algorithm. In this context, the use of the AL paradigm [1]-[8] has been recently introduced in the remote sensing literature as an effective tool for defining or enriching the training set in an interactive and iterative way. Considering applications where large geographical areas have to be classified (eventually considering several images), or where the land-cover map has to be (regularly) updated given new remote sensing images acquired on the same geographical area, the problem of defining adequate training sets becomes more and more important.

In this paper, we assume that a remote sensing image and the related reference labeled samples are available from a previous analysis and a land-cover map can be obtained according to a supervised classification technique. Our goal is to classify another image acquired on another geographical area with similar characteristics and the same land-cover classes. In such a scenario, it is very important to exploit the already available information from the first image in order to reduce the costs associated to the classification of the second one. From a machine learning perspective, this class of problems can be modeled in the framework of transfer learning, and in particular of DA, which goal is to transfer the information learned by the classification system from a source domain (associated with the first image) to a target domain (associated to the second image) [9]-[10]. In such problems, it is usually assumed that source and target domains have similar characteristics [9], i.e., they share the same set of information classes and the related class distributions are correlated (but not the same). Thus, the technique for addressing DA problems has to cope with the spatial/temporal variability of the spectral signatures of the land-cover classes in order to adapt the model of the classifier from source to target domain.

In the context of remote sensing literature, DA problems have been mainly addressed with semisupervised techniques [11]-[15], which exploit the labeled samples from the source domain and the unlabeled samples from target domain in order to derive a classification rule suitable for the target domain. In such problems it is assumed that labeled samples are available only for the source domain, but not for the target domain. Here, we suppose that some samples (as little as possible) from the target domain can be labeled by the user and added to the existing training set (defined on the source domain) in order to adapt the classifier to the target domain. Few papers addressed DA problems under this assumption [16]-[17], nonetheless they can deal only with a particular type of data-set shift between source and target domain (i.e., covariate-shift problems) and therefore cannot be generally adopted to cope with the variability of the spectral signatures of the land-cover classes in different remote sensing images.

In this paper, we propose a novel AL technique for addressing DA problems. The aim of the proposed technique is to accurately classify a second image (target domain), exploiting the information of a first image with reference labeled samples while requiring the minimum number of new-labeled samples from the second image. The proposed technique is based on the identification of the most informative samples of the target domain to be added to the training set, and the removal from the training set of the samples of the source domain, which are not descriptive of the target domain class distributions. The main novelties of this paper are represented by: 1) the introduction of a query function devoted to remove misleading samples from the source domain (called *q*-); and 2) an automatic stopping criterion for the iterative algorithm that does not need labeled samples from the target domain.

The outline of this paper is the following. In the next section, the theoretical background of DA and AL is presented. In Section III, a survey on related works is given. Section IV presents the proposed DA technique based on AL. In Section V, the adopted data sets and the design of experiments are described, and in section VI the obtained results are reported and discussed. Finally, Section VII draws the conclusions of this paper.

II. BACKGROUND ON DOMAIN ADAPTATION AND ACTIVE LEARNING

This section presents the background on the theoretical concepts on which this paper is based, i.e., transfer learning and DA problems as well as the AL approach. DA models different learning problems, while AL refers to the adopted strategy to address these kinds of problems.

A) Transfer learning and domain adaptation problems

Let us introduce some definitions and a general notation for presenting different types of learning to solve different kinds of classification problems. Two main families of learning can be used for training a classifier: supervised learning methods (when labeled training samples are given) and unsupervised learning methods (when labeled training samples are not available). Semisupervised learning is between supervised and unsupervised learning, i.e., both labeled and unlabeled samples are available and are jointly exploited by the classification algorithm [18].

With respect to the classification problems and settings, let us introduce the following concepts. A *domain D* consists of two components: a feature space $\mathcal{X}$ and a marginal probability distribution $P(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$ [19]. Given a domain $D = \{\mathcal{X}, P(\mathbf{x})\}$, a *learning task* consists of two components: a set of information classes Ω and an objective predictive function $f(\mathbf{x})$, which should be learned from the available data. From a probabilistic viewpoint $f(\mathbf{x})$ can be written as $P(y|\mathbf{x})$. Thus, for a given domain and learning task, the classification problem under investigation is governed by the distribution $P(\mathbf{x}, y) = P(\mathbf{x})P(y|\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$, $y \in \Omega$. In the supervised setting a training set $T\{(\mathbf{x}_j, y_j)\}_j$, $\mathbf{x}_j \in \mathcal{X}$, $y_j \in \Omega$ is available and classification algorithms are designed assuming that the predictive function $f(\mathbf{x})$ can be correctly estimated from the available data, i.e., $\hat{f}(\mathbf{x}) \approx f(\mathbf{x})$, and it is possible to obtain high classification accuracies over unseen test samples drawn from the same domain.

However, in many operational applications the available prior information is not sufficient to define a training set representative of the distribution to which the trained model should be applied, or the samples to be classified may be drown from a different domain. These kinds of problems can be addressed according to transfer learning (or knowledge transfer) methods. Transfer learning refers to the problem of retaining and applying the knowledge available for one or more domains, tasks, or distributions to efficiently develop an effective hypothesis for a new task, domain, or distribution. Instead of involving generalization across problem instances, transfer learning emphasizes the transfer of knowledge across tasks, domains, and distributions that are similar but not the same. Several types of

transfer learning problems have been addressed in the machine learning and data mining literature. Learning under *sample selection bias* is an important transfer learning problem, which occurs when unlabeled test data are drawn from the same domain D of the training data, but the estimated distribution $\hat{P}(\mathbf{x}, y) = \hat{P}(\mathbf{x})\hat{P}(y|\mathbf{x})$ does not correctly model the true underlying distribution that governs D since the number (or the quality) of available training samples is not sufficient for an adequate learning of the classifier. The small amount of labeled data generally yield to a poor estimation $\hat{P}(\mathbf{x})$ of the marginal distribution $P(\mathbf{x})$. Moreover, if the few available training data do not represent the general target population and introduce a bias in the estimated class prior distribution, this may cause a poor estimation of the conditional distribution, i.e., $\hat{P}(y|\mathbf{x}) \neq P(y|\mathbf{x})$. On the one hand, if both $\hat{P}(\mathbf{x}) \neq P(\mathbf{x})$ and $\hat{P}(y|\mathbf{x}) \neq P(y|\mathbf{x})$ the problem is referred to as *sample selection bias* [20], [21], [22]. On the other hand, the particular case where the true and estimated distributions are assumed to differ only via $\hat{P}(\mathbf{x}) \neq P(\mathbf{x})$ while $\hat{P}(y|\mathbf{x}) \approx P(y|\mathbf{x})$ is denoted by *covariate shift* [23], [24]. The aforementioned problems are very common in the classification of remote sensing images, where the available training sets are typically small and the samples are often represented in high dimensional feature spaces. This problem is particularly important in the classification of hyperspectral data, which are characterized by hundreds of spectral channels, or in very high resolution images (VHR), which require the extraction of several features to model the geometrical and textural properties of the scene. Semisupervised learning techniques have proved to be quite effective in addressing this kind of problems, improving the estimation of the class distributions exploiting also unlabeled samples (in addition to the labeled ones).

In this paper, we focus on DA problems, where test patterns $TS_t = \{x_j^t\}_j$, are drawn from a target domain D_t different from the source domain D_s of training samples $T_s\{(\mathbf{x}_j^s, y_j^s)\}_j$, $\mathbf{x}_j^s \in \mathcal{X}_s$, $y_j^s \in \Omega$. In this context, let $P^s(\mathbf{x}, y) = P^s(\mathbf{x})P^s(y|\mathbf{x})$ and $P^t(\mathbf{x}, y) = P^t(\mathbf{x})P^t(y|\mathbf{x})$ be the true underlying distributions for the source and target domains, respectively. We assume here that both the marginal distributions and the conditional distributions of the classes may differ in the two domains, i.e., $P^t(\mathbf{x}) \neq P^s(\mathbf{x})$ and $P^t(y|\mathbf{x}) \neq P^s(y|\mathbf{x})$. This is the most general case that we can expect by considering the distributions of the classes associated to two images acquired in different geographical areas. The key idea is to infer a good approximation of $P^t(\mathbf{x}, y)$ by exploiting $P^s(\mathbf{x}, y)$. If $P^t(y|\mathbf{x})$ is different from $P^s(y|\mathbf{x})$, DA is necessary. In remote sensing applications, this happen when source and target domains are associated to two different remote sensing images acquired on different areas with similar characteristics and same land-cover classes (or when the two images are acquired on the same

geographical area at two distinct times). We can here distinguish two varieties of DA problems that have been addressed in the literature [25]: 1) the semisupervised scenario, and 2) the supervised one. The semisupervised scenario models the case where a training set $T_s\{(\mathbf{x}_j^s, y_j^s)\}_j$, $\mathbf{x}_j^s \in \mathcal{X}_s$, $y_j^s \in \Omega$ is available for the source domain (associate to the first image), while only unlabeled data $U_t = \{x_j^t\}_j$ are available for the target domain (associated to the second image). In the supervised DA case a small amount of labeled samples are available also for the target domain $T_t\{(\mathbf{x}_j^s, y_j^s)\}_j$, $\mathbf{x}_j^s \in \mathcal{X}_s$, $y_j^s \in \Omega$ or part of the available unlabeled samples may be selected for labeling.

In the framework of domain adaptation, most of the learning methods are inspired by the idea that, although different, the two considered domains are correlated. In particular, it is intuitive to observe that considering the data of both source and target (unlabeled for the semisupervised case or labeled for the supervised one) domains in the training phase could improve the performances with respect to ignore one of the two information sources.

B) Active learning

AL is an approach to iteratively select the most informative samples for defining a training set by exploiting the classification rule [1]-[8]. To precisely describe the workflow of a general AL process, let us model it as a quintuple (G, Q, S, T, U) [26]. *G* is a supervised classifier, which is trained with the training set *T*. *Q* is a query function used to select the most informative unlabeled samples from a pool *U* of unlabeled samples on the basis of the current classification results. *S* is a supervisor who can assign the true class label to any unlabeled sample of *U* (e.g., a human expert). The AL process is an iterative process, where the supervisor *S* interacts with the system by labeling the most informative samples selected by the query function *Q* at each iteration. At the first stage, an initial training set $T^{(0)}$ of few labeled samples is required for the training of the classifier *G*. After initialization, the query function *Q* is used to select a set of samples from the pool *U* and the supervisor *S* assigns them the true class label. Then, these new labeled samples are included into $T^{(i)}$ (where *i* refers to the iteration number) and the classifier *G* is retrained using the updated training set. The closed loop of querying and retraining continues until a stop criterion is satisfied. Algorithm 1 gives a description of a general AL process.

Algorithm 1: Active Learning procedure

1. Train the classifier G with the initial training set T
2. Classify the unlabeled samples of the pool U

Repeat

3. Query a set of samples (with query function Q) from the pool U
4. A label is assigned to the queried samples by the supervisor S
5. Add the new labeled samples to the training set T
6. Retrain the classifier

Until a stopping criterion is satisfied.

In the classification of single images (i.e., single domain), AL techniques are very effective for defining a training set that can correctly describe the addressed classification problem according to the considered classification model. This is possible at the expense of conducting the labeling process iteratively, which may lead to increase or decrease the overhead costs depending on both the adopted labeling procedure (i.e., photointerpretation or *in situ* surveys) and the considered application. The AL approach has been applied to different classification context to prevent or address problems of learning under *sample selection bias* or *covariate shift*. Thus, AL can be considered alternative to semisupervised learning in addressing these kinds of problems. In this paper we will adopt an AL strategy to address supervised DA problems.

III. RELATED WORKS

In the last years, there has been a growing interest in developing both DA and AL methods, in different application fields. In this section we review the state of the art on: 1) DA methods; and 2) on AL and its use to address DA problems.

A) Domain Adaptation methods

In the context of remote sensing, DA techniques have been mainly studied in the semisupervised setting (i.e., without labeled target-domain data) for addressing the problem of automatic updating of land-cover maps [11]-[14]. In [11], a DA approach is proposed that is able to update the parameters of an already trained parametric maximum-likelihood (ML) classifier on the basis of the distribution of a new image for which no labeled samples are available. In [12], in order to take into account the temporal correlation between images acquired on the same area at different times, the ML-based DA approach is reformulated in the framework of the Bayesian rule for cascade classification (i.e., the classification

process is performed by jointly considering information contained in the source and target domains). The basic idea in both approaches is modeling the observed spaces by a mixture of distributions whose components are estimated by using unlabeled target data according to the Expectation Maximization (EM) algorithm with finite Gaussian Mixture Models. In [13], DA approaches based on a multiple-classifier system and a multiple-cascade-classifier system (MCCS) have been defined, respectively. In [14], the proposed MCCS architecture is composed of an ensemble of classifiers developed in the framework of cascade classification, which is integrated in a multiple-classifier architecture. A DA technique based on the use of a binary hierarchical classifier for the analysis of hyperspectral images is proposed in [27]. Both the semisupervised and supervised DA cases are considered. The authors in [28] proposed a semisupervised adaptation technique based on manifold regularization. In [15], a semisupervised SVM classifier based on the combination of clustering and the mean map kernel is proposed. In [29], a framework called Gaussian process maximum likelihood for spatially adaptive classification of hyperspectral data is proposed. This framework provides a way to model the spatial variations of the hyperspectral images characterizing each land cover class at a given location by a multivariate Gaussian distribution with parameters adapted for that location. In [9], a technique to solve semisupervised DA problems based on SVM is proposed. Such technique is designed to adapt an SVM classifier trained with source-domain samples to the target domain, exploiting both labeled source-domain samples and unlabeled target-domain samples. This is obtained with an iterative procedure where target-domain samples that are informative for the classifier (i.e., they are likely to become support vectors at the next iteration) and associated to the highest confidence in the correct label are included in the training set for re-training the SVM classifier at the next iteration. At the same time, source-domain samples are gradually erased in order to obtain a final classification rule defined only on the basis of target-domain samples.

Supervised DA problems have been addressed mainly in the context of text and natural language processing (NLP). In [10], a framework to address DA problems using the conditional expectation maximization algorithm is proposed, which is a variant of the standard EM algorithm for discriminative models. In [30], the authors propose a general instance weighting framework for DA problems that implements several adaptation heuristics: removing misleading training samples in the source domain, assigning more weights to labeled target patterns than labeled source patterns, and augmenting training samples with target samples with predicted labels. In [31], the authors propose a transfer learning framework called TrAdaBoost, which extends the standard boosting-based learning algorithms. Such a

method aims at boosting the accuracy of a weak learner by adjusting the weights of training samples, i.e., the weights of misclassified samples belonging to the target domain are increased, whereas the weights of misclassified samples belonging to the source domain are decreased. In this way the impact of source domain samples that are wrongly predicted due to distribution changes is weakened.

A) Active Learning methods in Domain Adaptation problems

AL methods have recently gained significant interest in the remote sensing community [1]-[8]. The use of AL for the classification of RS images was first introduced in [1]. The technique presented in such a paper is based on the selection of the most uncertain sample for each binary SVM in a One-Against-All (OAA) multiclass architecture. In [2], an AL technique is presented, which selects the unlabeled sample that maximizes the information gain between the a posteriori probability distribution estimated from the current training set and the training set obtained by including that sample into it. The information gain is measured by the Kullback–Leibler (KL) divergence. In [3], two AL techniques are proposed. The first technique is margin sampling by closest support vector (MS-cSV), which considers the smallest distance of the unlabeled samples to the *n* hyperplanes as the uncertainty value. The most uncertain samples that do not share the closest SV are added to the training set at each iteration. The second technique, called entropy query-by bagging (EQB), is a classifier independent approach based on the selection of unlabeled samples according to the maximum disagreement between a committee of classifiers. In [4], different batch-mode AL techniques for the classification with SVM are investigated. The investigated techniques exploit different query functions, which are based on both uncertainty and diversity criteria. Moreover, both a query function (which is based on a kernel-clustering technique for assessing the diversity of samples) and a strategy for selecting the most informative representative sample from each cluster are proposed. The AL method presented in [5] is based on the cluster assumption and considers the 1-D output space of the SVM classifier to identify the most uncertain samples lying in the low-density region. This technique is designed to address critical problems where the available initial training set is significantly biased. In the AL method proposed in [6], the unlabeled samples are queried according to the output of an SVM classifier trained to discriminate between significant and nonsignificant samples. Support vectors are considered as significant samples, whereas all the other samples are defined as nonsignificant. In [7], a co-regularization framework for AL is proposed for hyperspectral image classification. The first regularizer exploits the intrinsic multi-view information embedded in the hyperspectral data, while the second one is based on the cluster assumption

and designed on a spatial or a spectral based manifold space. An AL technique for defining effective multitemporal training sets to be used for the supervised detection of land-cover transitions in a pair of remote sensing images acquired on the same area at different times in proposed in [8]. This technique is developed in the framework of the Bayes' rule for compound classification and aims at selecting the pair of spatially aligned unlabeled pixels in the two images that are classified with the maximum uncertainty.

The use of AL for addressing supervised DA problems was introduced very recently in the field of NLP in [32], [33]. In [32], the authors adopt an AL strategy to DA for word sense disambiguation, in order to select the samples of the target domain to be labeled and added to the training set. This represents a preliminary work on the use of AL for DA problems in which the classifier is first trained using source-domain labeled samples, and then at each iteration, the samples of the target domain are selected by the query function one by one and added to the original training set. In [33], two different algorithms for DA with online AL are proposed. The first algorithm, called active online DA (AODA) works is two steps: 1) the classifier is first trained using only the source-domain data, then 2) the classification rule is iteratively updated in an online fashion actively requiring labeled samples from target domain. The second presented algorithm, called domain-separator based AODA, is a variant of the first one that aims at better exploiting the relatedness between source and target domain. This is obtained by ignoring the labeled samples from the target domain that lie in the same portion of the feature space of the source-domain samples (thus assuming that $P^t(y|\mathbf{x}) = P^s(y|\mathbf{x})$). In the context of remote sensing, Jun *et al.* proposed in [16] a method based on the same philosophy as in [31] to the classification of hyperspectral data, modified for AL strategy. The aforementioned work proposed a methodology where the weight updating rules were determined heuristically and the AL algorithm is based on the KL-max method [2]. In [17], an active learning technique to address data set shift problems under covariate shift is proposed. Two criteria based on uncertainty and clustering of the data space are considered to perform active selection. The use of a clustering-based selection strategy allows the user to discover new classes in the case they have been omitted in the initial training set.

IV. PROPOSED DOMAIN ADAPTATION TECHNIQUE BASED ON ACTIVE LEARNING

In this paper, we focus on supervised DA problems, where a limited number of target-domain samples can be queried by the automatic system for manual labeling by the user. The main idea of the

proposed approach is to iteratively labeling and adding to the training set the minimum number of the most informative samples from the target domain, while removing the source-domain samples that do not fit with the distributions of the classes in the target domain. This idea is partially inspired by the DASVM technique [15], where target-domain samples are gradually included in the training set and source-domain sample are removed at the same time by a semisupervised learning technique. However, here we extend the DA approach to the supervised case in the framework of Bayes decision theory, which results in important differences in the strategy for selecting and removing samples from the domains. We use an AL procedure that interacts with the user, which is requested to annotate the samples selected by the query function on the basis of their uncertainty. The labeled samples are then added to the current training set for retraining the classifier. In this way, the classification system exploits already available information (i.e., the labeled samples of the source domain) in order to drive the user in the annotation of the most informative samples of the target image, minimizing the number of samples to be labeled. Depending on the correlation between the two classification problems, the number of training samples that should be used for classifying the new image can be strongly reduced.

The developed technique is based on the ML classifier, which is a simple yet appropriate technique for the classification of multispectral and hyperspectral images that finds application in many real world problems. Given a classification problem characterized by the set of information classes $\Omega = \{\omega_1, \omega_2, \dots, \omega_C\}$ and the class-conditional densities $p(\mathbf{x}|\omega_i)$, $i = 1, \dots, C$, estimated from the available training samples, the classification label $y_t \in \Omega$ of a test sample $\mathbf{x}_t$ is obtained using the following rule:

$$y_t = \underset{\omega_n \in \Omega}{\operatorname{argmax}} \{p^{(i)}(\mathbf{x}_t|\omega_n)\}. \quad (1)$$

According to previous studies (e.g., [2], [11], [16], [34]-[36]), in our implementation we modeled the class-conditional densities with Gaussian distributions. The estimation of the covariance matrices of the classes is carried out according to the "leave-one-out covariance matrix estimate" (LOOC) proposed in [35]. This technique is specifically designed to give robust estimates of the second-order class statistics even when the available training samples are few compared to the number of features. This increases the robustness of the traditional ML classifier, allowing the user to start the AL process with a limited quantity of training samples. The effectiveness of such a technique in the classification of high dimensional data sets has been reported in [35], [36]. Nevertheless, the proposed approach can be generalized to other classifiers.

The proposed strategy for addressing DA problems with AL is based on the definition of two query functions, namely: 1) "query+" ($q+$), which is devoted to select the most informative samples from target domain on the basis of their uncertainty, and 2) "query-" ($q-$), whose aim is to remove from the current training set source-domain samples which are not representative of the target-domain problem. Exploiting these two query functions, the proposed technique progressively introduces new target-domain samples in the training set, while removing source-domain samples. In this way, the classification system iteratively adapts itself to the classification of the target-domain problem, asking the user to label a reduced number of new samples. It is worth noting that $q+$ is a standard concept for AL techniques, whereas $q-$ is a novel concept introduced in this paper to address DA problems. In the following subsections, part A and B describe the proposed $q+$ and $q-$ functions, respectively; part C describes their combination for the proposed DA technique; and part D addresses the problem of defining a stop criterion for the iterative procedure.

A) Query+: The aim of $q+$ is to select the batch X^+ of the most informative samples from the pool U of unlabeled samples, which are taken from the target domain D_t. Once selected, such samples are labeled by the user, and added to the training set T. In our implementation we adopted a *breaking ties* approach [37], which is inspired to the *multiclass-level uncertainty* with *difference function* presented in [4] for classification with SVMs. The sample $\mathbf{x}^+$ minimizing the difference between the largest and the second largest class-conditional densities is selected according to the following equation:

$$\mathbf{x}^+ = \underset{\mathbf{x}\in U}{\operatorname{argmin}}\left\{p^{(i)}(\mathbf{x}|\omega_{max1}) - p^{(i)}(\mathbf{x}|\omega_{max2})\right\}, \tag{2}$$

where:

$$\omega_{max1} = \underset{\omega_n\in\Omega}{\operatorname{argmax}}\left\{p^{(i)}(\mathbf{x}|\omega_n)\right\}, \tag{3}$$

$$\omega_{max2} = \underset{\omega_m\in\Omega\setminus\{\omega_{max1}\}}{\operatorname{argmax}}\left\{p^{(i)}(\mathbf{x}|\omega_m)\right\} \tag{4}$$

The probability density function $p^{(i)}(\mathbf{x}|\omega_n)$ of pattern $\mathbf{x}$ conditional to class ω_n is computed using the Gaussian model estimated from the training set $T^{(i)}$ (available at iteration i). This approach results in the selection of the sample $\mathbf{x}^+$ associated to the highest uncertainty between the two most likely information classes. This procedure is repeated until the desired number of samples h^+ (defined by the user) has been selected. The selected samples are then removed from U, labeled by the user and added to the training set $T^{(i+1)}$ for the training of the classifier at the next iteration.

B) Query-: The aim of $q-$ is to remove from the source-domain training set T_s the labeled samples that do not fit with the distribution of the classes in the target domain and therefore are expected to

reduce the classification accuracy in such a domain. In order to identify a set X^- of these samples at each iteration, we consider the difference between the value of the class-conditional densities computed using only source-domain samples (i.e., using $T_s = T^{(0)}$) and those obtained using the training set at i-th iteration $T^{(i)}$ for each sample $\mathbf{x} \in T_s$. If such a difference is small, it means that the distribution of the class ω_l (to which $\mathbf{x}$ belongs to) at the iteration i does not change significantly with respect to the distribution computed with the original training set. On the contrary, if this difference is high, it means that the distribution of the class ω_l is shifting from source to target domain and the sample $\mathbf{x}$ is no more representative of the target-domain distribution. Thus, q- selects the sample $\mathbf{x}^-$ that maximizes such a difference as follows:

$$\mathbf{x}^- = \underset{\mathbf{x} \in T^{(0)}}{\operatorname{argmax}} \left\{ p^{(0)}(\mathbf{x}|\omega_l) - p^{(i)}(\mathbf{x}|\omega_l) \right\}. \tag{5}$$

In order to select a batch set X^- of h^- samples to be removed from the training set at each iteration, the aforementioned procedure is repeated h^- times.

Fig. 1 shows a qualitative example of the choice of the sample $\mathbf{x}^-$ by the q- function. The figure depicts the case of a two-classes classification problem defined in a one-dimensional feature space. The distribution of class ω_1 is shifting significantly from the source-domain distribution to the one estimated at the i-th iteration, while class ω_2 remains more stable**.** Thus, the sample $\mathbf{x}^- \in \omega_1$ associated to the maximum difference between the distribution of the class at the initial iteration and iteration i is selected by q-.

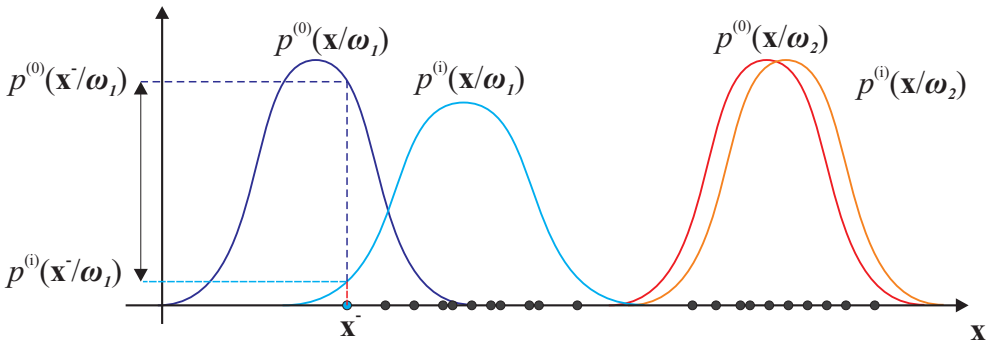


Fig. 1 – Qualitative example in a one-dimensional feature space showing the selection of the source-domain sample $\mathbf{x}^-$ that is not representative of the target-domain distribution.

A possible problem that might be caused by q- is related to the removal of too many samples from a particular class leading to a poor estimation of its distribution. Moreover, if the available samples for a class are too few (or not representative of the distribution), this leads to the estimation of a singular

covariance matrix, preventing the classifier to derive a reasonable classification rule. This occurs when a particular class distribution is significantly changing from one iteration to the next one, but few samples are available for its estimation. Thus, we implemented a mechanism to prevent the q- to remove samples from the source-domain training set if the number of samples of a class is lower than a threshold. This technique assures the classifier to have a minimum number of samples for each class at each iteration.

C) Combination of Query+ and Query-: The proposed DA technique aims at adapting the classification rule from source to target domain by exploiting both q+ and q- at the same time. Important user-defined parameters are h^+ and h^-, and in particular the ratio $\alpha = h^-/h^+$. The optimum value of the parameter α depends on both the size of T_s and the correlation between source and target domains. In the case the two domains are similar, a low value of h^- is sufficient, because it is expected that only few samples from the source domain should be removed from the training set. If $P^t(y|\boldsymbol{x}) \approx P^s(y|\boldsymbol{x})$ the value of h^- should tend to zero. Conversely, if the two domains are significantly different, and in particular if $P^t(y|\boldsymbol{x})$ is significantly different from $P^s(y|\boldsymbol{x})$, the value of h^- should be high and α should be set to a high value. In this case, the value of h^- depends also on the size of T_s: the higher is the size of T_s, the higher the value of h^- should be set. The proposed procedure is described in Algorithm 2.

The proposed technique can converge to solutions where all the source-domain samples are removed from the training set and the classification rule depends only on target domain samples. However, we expect that depending on the correlation between the two domains, convergence may be reached before, so that the final classification rule depends on both source and target-domain samples (thus reducing the number of samples to be labeled for the image to be classified). From a theoretical point of view, this approach allows one to address different DA problems associated with different statistical distributions of the classes in the two domains. In particular, it allows one to address problems where both the marginal distributions and the conditional distributions are different on the two domains, i.e., $P^t(\mathbf{x}) \neq P^s(\mathbf{x})$ and $P^t(y|\mathbf{x}) \neq P^s(y|\mathbf{x})$. This is the type of dataset-shift that we expect in general to have by considering two images characterized by spectral signatures of the different classes that can change because of several physical factors (e.g., illumination conditions, soil moisture, topography). Such cases are closely related to *sample selection bias* problems. In these situations, the optimal classification rule is likely to change significantly from source to target domain and, furthermore, misleading source-domain samples must be removed from the training set to accurately classify the target domain. Conversely, in [17], it is assumed that source and target domains differ only in the

marginal distributions (i.e., $P^t(\mathbf{x}) \neq P^s(\mathbf{x})$), whereas the conditional distributions remain substantially unchanged $P^t(y|\mathbf{x}) \approx P^s(y|\mathbf{x})$. This problem is closely related to the *covariate shift* problem. However, this assumption does not model the general type of data-set shift that we can expect in the considered application. The weighing mechanism adopted in [16] allows one to decrease the weights of misleading source-domain samples. Nevertheless, according to the proposed heuristic, such weights slowly tend to zero when the size of the target-domain samples included in the training set increases and never reach this value. Thus, misleading source-domain samples will still negatively affect the classification rule.

Fig. 2 show two qualitative examples associated to the two aforementioned different DA problems. In the first example (fig. 2a), the distribution of the classes in the two domains differs mainly because of a different sampling, i.e., $P^t(\mathbf{x}) \neq P^s(\mathbf{x})$, but the conditional distributions remain similar, i.e., $P^t(y|\mathbf{x}) \approx P^s(y|\mathbf{x})$. In the second case (fig. 2b) the two distributions are characterized by both $P^t(\mathbf{x}) \neq P^s(\mathbf{x})$ and $P^t(y|\mathbf{x}) \neq P^s(y|\mathbf{x})$.

Algorithm 2: Proposed DA technique based on AL

Input arguments:
h^+: number of samples of the target domain to be queried at each iteration;
h^-: number of samples of the source domain to be removed from the training set at each iteration;
T_s : training set defined on the source domain;
U : pool of unlabeled samples of the target domain which should be labeled by S when queried;

1. Train the ML classifier with the initial training set $T^{(0)} = T_s$
2. Classify the unlabeled samples of the pool U

Repeat

3. Select the set X^+ of h^+ samples from U using $q+$
4. Select the set X^- of h^- samples from $T^{(i)}$ using $q-$
5. The supervisor S assigns a label to the samples in X^+
6. Update the training set $T^{(i+1)} = \{T^{(i)}/X^-\} \cup X^+$
7. Retrain the classifier with $T^{(i+1)}$

Until a stopping criterion is satisfied.

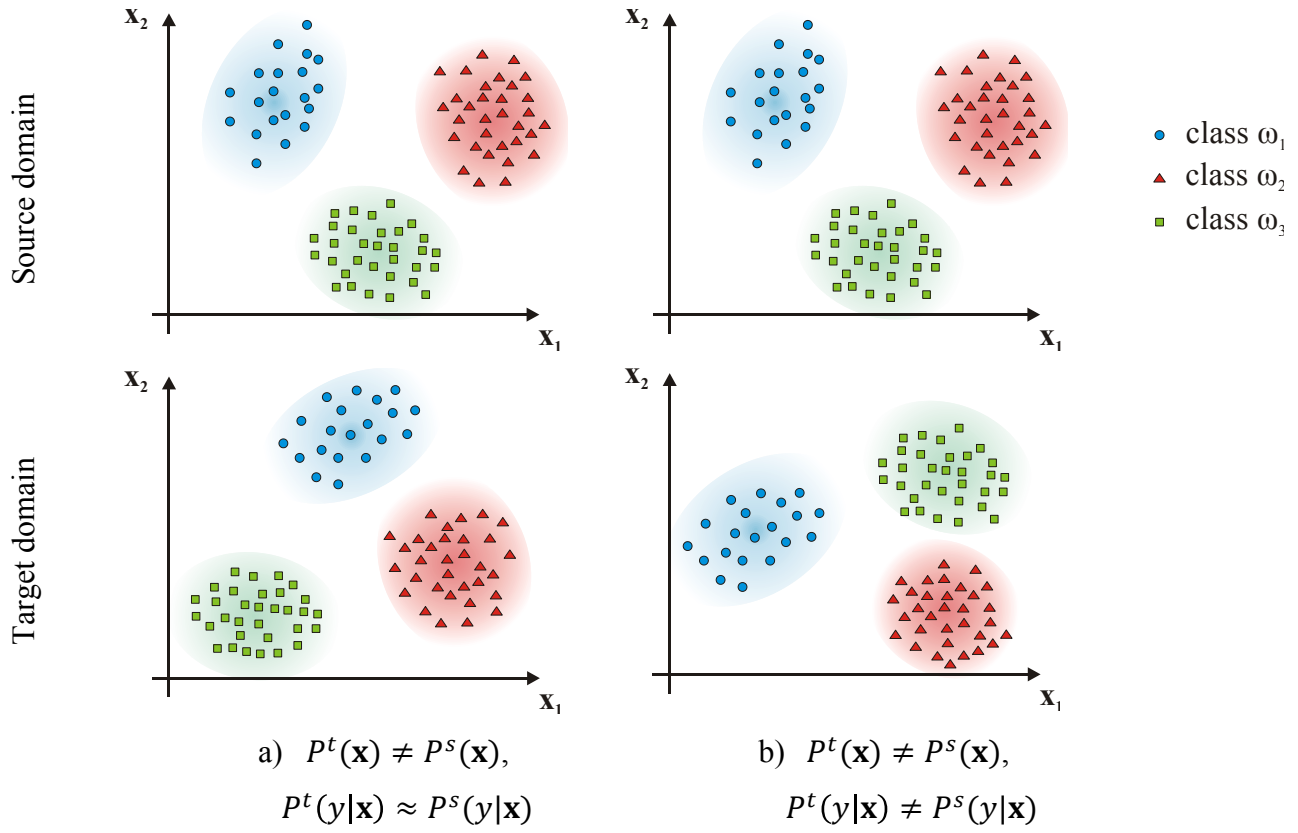


Fig. 2 – Qualitative examples of different DA problems. a) DA problem characterized by $P^t(\mathbf{x}) \neq P^s(\mathbf{x})$, but $P^t(y|\mathbf{x}) \approx P^s(y|\mathbf{x})$. This case is closely related to *covariate shift* problems. b) DA problem characterized by $P^t(\mathbf{x}) \neq P^s(\mathbf{x})$ and $P^t(y|\mathbf{x}) \neq P^s(y|\mathbf{x})$. This case is related to *sample selection bias* problems.

D) Stopping criterion: An important issue to be considered with the proposed AL procedure is to define an effective stopping criterion for the iterative procedure, which should be able to automatically detect the convergence of the algorithm. If a test set defined on the target domain is available, the AL procedure can be iterated until the classification accuracy computed on such a test set reaches a (relatively) stable value. At the saturation point, the addition of new-labeled samples cannot significantly increase the classification accuracy and thus the algorithm is in convergence. However, in operative classification problems, it is not realistic to assume that a test set is available for the target domain. In this case, we propose to identify the convergence of the AL technique by considering only the labeled samples $T^{(i)}$ available at the *i*-th iteration and the initial training set $T^{(0)}$ defined on the source domain. Computing a statistical distance measure between the class distributions of $T^{(i)}$ and $T^{(0)}$ it is possible to obtain information about how the class-conditional densities at the *i*-th iteration have moved away from those of the source domain. When such a distance reaches saturation to a stable value, we can assume convergence and thus stop the iterative algorithm. In our implementation we considered

the Bhattacharyya distance $B_n(i)$ [37], which can be adopted to evaluate the statistical distance between the class-conditional densities $p^{(i)}(\mathbf{x}|\omega_n)$ and $p^{(0)}(\mathbf{x}|\omega_n)$ for all classes $\omega_n \in \Omega$ at a given iteration i**.** The general model-free definition of such a distance for a generic class $\omega_n \in \Omega$ is as follows:

$$B_n(i) = -\ln\left\{\int_x \sqrt{p^{(i)}(\mathbf{x}|\omega_n)p^{(0)}(\mathbf{x}|\omega_n)}\right\}. \tag{6}$$

Adopting a Gaussian model for the class-conditional densities, the Bhattacharyya distance $B_n(i)$ can be rewritten as:

$$B_n(i) = \frac{1}{8}\left(\mu_n^{(i)} - \mu_n^{(0)}\right)^T \left(\frac{\Sigma_n^{(i)} + \Sigma_n^{(0)}}{2}\right)^{-1} \left(\mu_n^{(i)} - \mu_n^{(0)}\right) + \frac{1}{2}\ln\left(\frac{\left|\frac{\Sigma_n^{(i)} + \Sigma_n^{(0)}}{2}\right|}{\sqrt{\left|\Sigma_n^{(i)}\right|\left|\Sigma_n^{(0)}\right|}}\right), \tag{7}$$

where $\mu_n^{(i)}$ and $\mu_n^{(0)}$ are the mean vectors of the class ω_n computed using $T^{(i)}$ and $T^{(0)}$ respectively, while $\Sigma_n^{(i)}$ and $\Sigma_n^{(0)}$ are the covariance matrices of the class ω_n obtained using $T^{(i)}$ and $T^{(0)}$, respectively. In order to find a unique saturation point for all the C considered classes (equally weighting the classes), the mean Bhattacharyya distance $B(i)$ can be calculated averaging $B_n(i)$ over all the classes $\omega_n \in \Omega$, i.e.:

$$B(i) = \frac{1}{C}\sum_{n=1}^{C} B_n(i) \tag{8}$$

If the user is more interested in some specific classes instead of the overall classification accuracy, $B(i)$ can be alternatively defined as a weighted average with user-defined weights. We can define a stopping criterion that interrupts the iterative procedure when $B(i)$ reaches saturation. The saturation point can be detected when the following inequality is satisfied:

$$B(i) - B(i-1) < \varepsilon, \tag{9}$$

where ε is defined by the user. If $B(i)$ presents a non-monotonic behaviour, a more robust method to detect the saturation point can be derived considering the average of $B(i)$ over s previous iterations, $h(i) = 1/(s+1)\sum_{r=0}^{s} B(i-r)$ and adopting the following alternative rule:

$$h(i) - h(i-s-1) < \varepsilon. \tag{10}$$

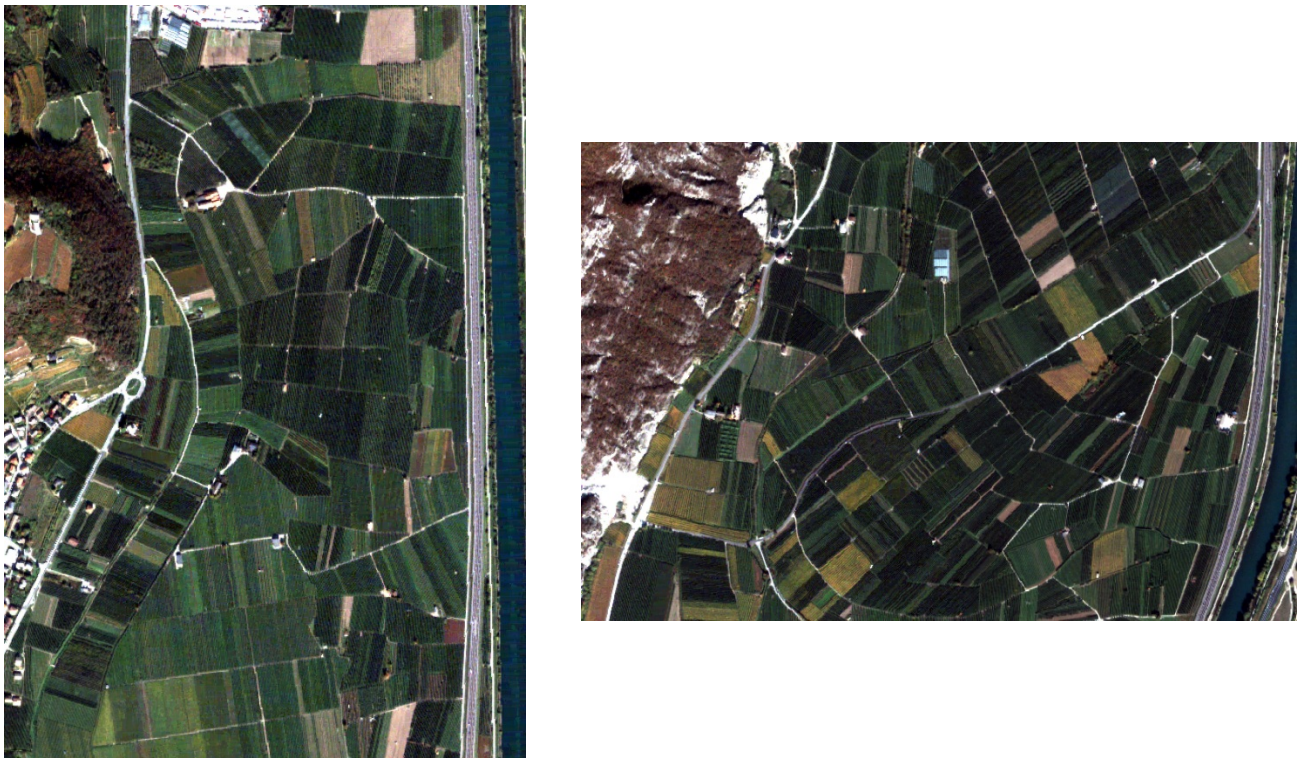
In this way we can control the AL in our DA problem also without test samples for the target domain (which is a common situation in operative scenarios).

V. DATA SET DESCRIPTION AND DESIGN OF EXPERIMENTS

We carried out different experiments in order to assess the effectiveness of the proposed technique and compare it with state-of-the-art techniques considering both a multispectral VHR and a hyperspectral data set. The description of the two data sets and the related design of the experiments are given below.

A) VHR Multispectral Data Set

The first data set is made up of two multispectral VHR images acquired by the Quickbird satellite over two rural areas in the city of Trento, Italy. The spatial resolution of the multispectral channels is 2.8 m, while the panchromatic band has a geometric resolution of 0.7 m. The first image (called here QB_1) consists of 2066 × 2983 pixels, while the size of the second image (QB_2 hereafter) is 3100 × 2066 pixels. True color compositions of the two images are shown in Fig. 3. Reference labeled samples are available for both images, corresponding to the following land-cover classes: 1) vineyard, 2) water, 3) agriculture fields, 4) forest, 5) apple tree, 6) urban area. Such labeled samples were collected through both photointerpretation and ground surveys. In our experiments, we used QB_1 as first image associated to the source domain and QB_2 as second image associated to the target domain. Relative to QB_1, a training set T_1 of 4427 samples and a test set TS_1 of 2176 samples are available. For the image QB_2, a training set T_2 and a test set TS_2 made up of 4668 and 15964 samples are available, respectively. Information about samples distributions on the different classes and sets is reported in TABLE I. For characterizing the discrepancy between the two domains, we computed the Bhattacharyya distance between the distributions of the classes in QB_1 and QB_2, considering all available labeled samples from the two images. The computed distances are reported in the last column of TABLE I. In order to give also a visual representation of the DA problem, Fig. 4 shows the distribution of the labeled samples considering bands 3 and 4 of the Quickbird images available for the source and the target domains, respectively. The two scatterplots show the distribution of the classes in the two domains, clearly highlighting the changes in the class distributions and the relationship between the source and target domains on the considered features. Please note that the distribution of the classes depicted in the figure are in agreement with the Bhattacharyya distances, highlighting that the classes “agriculture fields” and “forest” are associated with the more significant shift, while the other classes are more stable in the considered feature space.

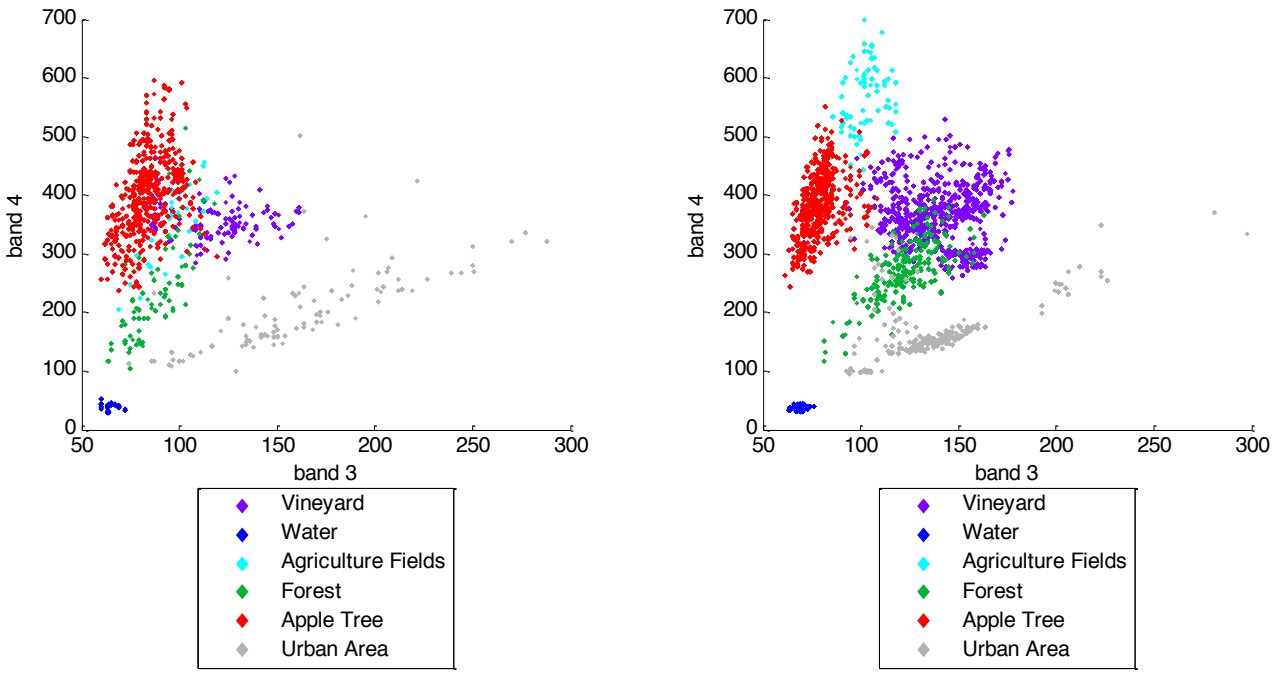


a) QB_1 b) QB_2

Fig. 3 – True color compositions of the two multispectral VHR images acquired by the Quickbird satellite. a) QB_1 image associated to the source domain. b) QB_2 image associated to the target domain.

TABLE I - NUMBER OF TRAINING (T_1 AND T_2) AND TEST (TS_1 AND TS_2) PATTERNS AVAILABLE IN THE TWO CONSIDERED IMAGES QB_1 AND QB_2. THE LAST COLUMN REPORTS THE BHATTACHARYYA DISTANCES BETWEEN THE DISTRIBUTIONS OF THE CLASSES IN QB_1AND QB_2 (VHR MULTISPECTRAL DATA SET).

Class	**Number of samples**				**Bhattacharyya distance**
	Image QB_1		**Image QB_2**		
	T_1	***TS_1***	***T_2***	***TS_2***	
Vineyard	658	314	848	6677	0.4318
Water	98	32	266	1180	0.6953
Agriculture Fields	105	45	260	620	2.8056
Forest	272	146	332	2434	1.6632
Apple Tree	3060	1523	2712	3273	0.1599
Urban Area	234	116	250	1780	0.6055
Total/Average	4427	2176	4668	15964	1.0602



b) b)

Fig. 4 – Distributions of the labeled samples on bands 3 and 4 of the Quickbird images. a) Source domain labeled samples (considering both training T_1 and test set TS_1). b) Target domain labeled samples (considering both training set T_2 and test set TS_2).

The experiments on the VHR data set were carried out in order to adapt the ML classifier trained on QB_1 to the classification of QB_2. We defined ten different initial training sets optimized for the classification of the source domain QB_1 (i.e., maximizing the accuracy on TS_1). Such initial training sets were obtained by applying an AL procedure based on the presented $q+$ using T_1 as pool and TS_1 for accuracy assessment. The size of the ten initial training sets is 965 samples. Such training sets were used for training the ML classifier at the first iteration in ten different trials. As a pool U for the AL process we considered T_2 (not considering the labels of the samples). We compared the results obtained by the proposed DA method with those yielded using: 1) only the $q+$ function, 2) random selection of samples from U, and 3) the $q+$ function combined with the sample weighting heuristic proposed in [16]. For comparison purposes, we also applied AL directly to QB_2, without considering the information available from QB_1. In this last case we randomly selected sets made up of 240 samples from T_2 and used the remaining samples as pool. Using less than 240 samples resulted in the impossibility to correctly

estimate the covariance matrices for all the considered classes (which can be singular or close to singular). We then ran the learning process on the target domain considering both the proposed $q+$ function and random sampling (passive learning). These two experiments aim at comparing the results of the proposed DA technique with those obtained by traditional techniques, which cannot exploit the information available from the source domain.

Finally, we run other two sets of experiments with the purpose of studying the impact of the initial training set size on the DA technique. We performed two trials of ten training sets made up of 485 and 1925 samples, representing approximately half and double size of the initial training sets size used in the first trial, respectively. Such training sets were obtained with the same procedure used for the first trial. Similar experiments were carried out using these initial training sets.

B) Hyperspectral Data Set

The second data set is a hyperspectral image that is used as a benchmark in the literature and consists of data acquired by the Hyperion sensor of the EO-1 satellite in an area of the Okavango Delta, Botswana. The considered image has a spatial resolution of 30 m over a 7.7 km strip in 145 bands. For greater details on this data set, we refer the reader to [2]. Reference labeled samples of 14 land-cover classes are available for two different and spatially disjoint areas, which are referred in the following as Area 1 and Area 2, representing two different geographical areas with the same set of land-cover classes characterized by slightly different distributions. The labeled samples taken from Area 1 were randomly partitioned into two sets T_1 and TS_1 and the samples of Area 2 were similarly partitioned into a training set T_2 and a test set TS_2, as in [34] (see TABLE II). From the original spectral bands we selected the subset of ten bands that maximized the discrimination capability among the classes according to the Jeffries-Matusita distance [37] as done in [34]. This step allowed us to remove both redundant and non discriminant bands from the hyperspectral data and to reduce the ratio between the number of features and the number of labeled samples per class (improving therefore the generalization capability of the classifier). Also for this data set we computed the Bhattacharyya distance between the distributions of the classes in Area 1 and Area 2, considering all available labeled samples from the two areas. The computed distances are reported in the last column of TABLE II. Please note that the classes “Hippo grass”, “Mixed mopane” and “Acacia grasslands” are associated to a significant shift, while the other classes are more stable.

The experiments on the hyperspectral data set were carried out in order to adapt the ML classifier trained on the Area 1 (considered as source domain) to the spatially separate Area 2 (considered as target domain). Also for the hyperspectral data set, ten different initial training sets made up of 707 samples optimized for the classification of Area 1 were selected as done for the multispectral VHR data set. These training sets were used for training the ML classifier at the first iteration in ten different trials. As pool U for the AL process we considered T_2, and TS_2 was used as test set to evaluate the classification accuracy on the target domain at each iteration. We compared the results obtained by the proposed DA method combining both $q+$ and $q-$ (using different values of α) with those obtained using only $q+$, a random selection of samples and the combination of our $q+$ function with the sample weighting heuristic proposed in [16]. We also applied AL directly to the target domain, without considering the information available from Area 1. In this case, we observed that selecting less than 270 samples from T_2 did not lead to a correct estimation of the covariance matrices for all the considered classes, and did not allow us to perform classification. Thus, we randomly selected 10 initial training sets of 270 samples from T_2 and used the remaining samples as pool. We ran the learning process on the target domain considering both the proposed $q+$ function and random sampling.

TABLE II - NUMBER OF TRAINING (T_1 AND T_2) AND TEST (TS_1 AND TS_2) PATTERNS AVAILABLE IN THE TWO SPATIALLY DISJOINT AREAS. THE LAST COLUMN REPORTS THE BHATTACHARYYA DISTANCE BETWEEN THE DISTRIBUTIONS OF THE CLASSES IN AREA 1 AND AREA 2 (HYPERSPECTRAL DATA SET).

Class	**Number of samples**				**Bhattacharyya distance**
	Area 1		**Area 2**		
	T_1	TS_1	T_2	TS_2	
Water	69	57	213	57	0.6452
Hippo grass	81	81	83	18	3.5419
Floodplain grasses1	83	75	199	52	1.8268
Floodplain grasses2	74	91	169	46	1.1583
Reeds1	80	88	219	50	0.8273
Riparian	102	109	221	48	0.3672
Firescar2	93	83	215	44	1.2217
Island interior	77	77	166	37	1.9668
Acacia woodlands	84	67	253	61	1.2441
Acacia shrublands	101	89	202	46	0.5472
Acacia grasslands	184	174	243	62	2.6810
Short mopane	68	85	154	27	0.8733
Mixed mopane	105	128	203	65	2.8093
Exposed soil	41	48	81	14	1.2196
Total/Average	1242	1252	2621	627	1.4950

VI. EXPERIMENTAL RESULTS

A) Results on the VHR Multispectral Data Set

Fig. 5 shows the classification accuracies on TS_2 (averaged over ten trials) obtained with the considered techniques versus the number of labeled samples of the target domain (selected from the pool U) added to the training set for learning the ML classifier. The reported curves are obtained fixing $h^+ = 2$ for all methods, and $h^- = 10$ for the proposed DA method. Since no diversity criteria are considered in the query functions, small h^+ were considered in order to avoid significant redundancy among the samples selected in the batch. Indeed, some preliminary trials (not reported in this paper) showed that values of $h^+ > 2$ result in lower classification accuracies. The optimum value of h^- was selected carrying out different trials with $h^- = 2, 4, \dots, 20$.

The obtained results clearly show that the proposed DA technique, which uses the combination of $q+$ and $q-$, led to significantly higher accuracies on the target domain compared to standard AL methods using only either the $q+$ function or a random selection. The proposed method significantly outperformed also the AL method based on the weighting heuristic presented in in [16]. This confirms the importance of the $q-$ query and the effectiveness of the proposed technique for addressing DA problems. Moreover, it is important to note that the proposed method reached the convergence with a significant smaller number of samples than those required by applying AL directly to the target image QB_2. Thus, the proposed DA strategy allows one to obtain accurate classification maps of the target domain, effectively exploiting the information available from source domain, and including relatively few labeled samples from the target domain.

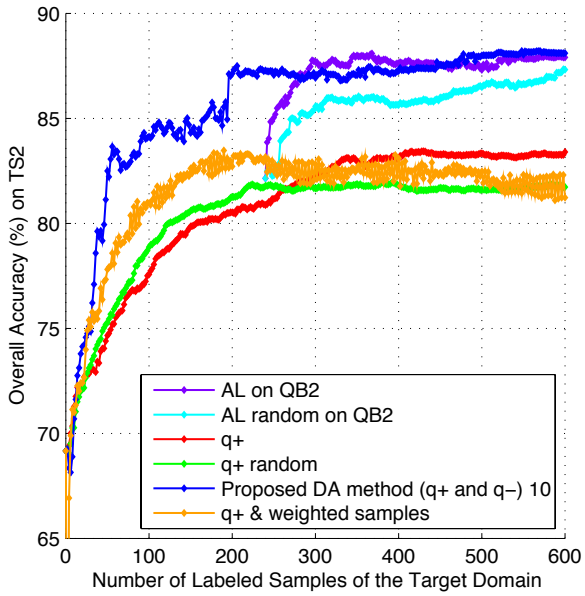


Fig. 5 – Overall classification accuracy (%) (averaged over ten trials) obtained on the test set TS_2 of the image QB_2 (target domain) versus the number of labeled samples of the target domain (selected from U) added to the training set by the different considered methods (multispectral VHR data set).

Considering the learning curves obtained with $q+$ and random selection, it might surprise that random selection preforms better than $q+$ up to 262 target-domain samples included in the training set. Nevertheless, this can be explained as follows. The $q+$ function selects the most informative samples of the target domain to be added to the training set on the basis of the estimated $p^{(i)}(\mathbf{x}|\omega_n)$ available at iteration i. However, if the estimated class-conditional densities do not correctly reflect the real densities on the target domain, the selection will be biased. In standard active learning, where the samples of the training set and pool are drawn from the same distribution (e.g., when AL is applied to the classification of a single image), it can be observed that in the first iterations, where the estimation of the real class-conditional densities is not precise yet, the $q+$ can lead to sub-optimal selections [5]. In the considered DA problem, where the training and pool samples are drawn from different distributions and domains, if the $q+$ function is not associated with the $q-$ function, such estimations are especially biased by the presence of source domain samples that are misleading for estimating the target domain distribution $P^t(\mathbf{x}|y)$. Thus, the estimations $p^{(i)}(\mathbf{x}|\omega_n)$ are biased and the $q+$ will concentrate the selection of samples in a region of the feature space that is not the most informative.

Regarding the results obtained using the weighting heuristic proposed in [16], we observe that this strategy converges to a classification accuracy significantly smaller than the one obtained with the proposed method. This behavior can be explained by the fact that misleading source-domain samples are associated to weights that slowly tend to zero when the number of the target-domain samples included in the training set increases, but never reach this value. Thus, such misleading samples affect the classification rule preventing the algorithm to adapt to the new classification problem defined on the target domain. Moreover, we observe that such a method leads to learning curves characterized by an oscillating behavior.

Fig. 6 shows the Bhattacharyya distances $B_n(i)$ and $B(i)$ between the class distributions computed using the training set $T^{(i)}$ and the initial training set $T^{(0)}$ (obtained using the proposed DA method at the *i*-th iteration) versus the number of new labeled samples of *U* considered in $T^{(i)}$. The rationale of this figure is to show how the distributions of the classes are moving away from the source domain (while they are getting closer to the target domain). From this plot it is clear that when about 200 labeled samples of the target domain are included in the training set the proposed method reached the convergence. This is in agreement with the plot of the accuracy on the test set TS_2 reported on Fig. 5. Using (10) to detect the saturation point of $B(i)$ (with s = 4 and $\varepsilon = 2e - 3$), the algorithm is stopped when 204 samples from target domain are added to the training set. Thus, we can conclude that considering only the labeled samples of $T^{(0)}$ and $T^{(i)}$, the proposed stop criterion is able to detect the convergence without the need for a test set defined on the target domain.

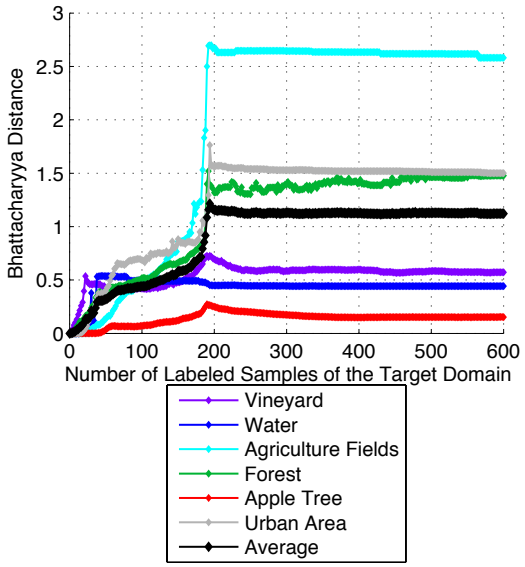


Fig. 6 – Bhattacharyya distances $B_n(i)$ and $B(i)$ between the class distributions computed on the initial training set $T^{(0)}$ and on the training set $T^{(i)}$ (obtained using the proposed DA method) versus the number of labeled samples of the target domain considered in $T^{(i)}$. (Multispectral VHR data set, proposed DA method).

TABLE III and TABLE IV report the classification results at different iterations in terms of overall accuracy, kappa coefficient and producer accuracies of the considered classes obtained with the proposed DA technique and with the method based on the combination of $q+$ and the weighing strategy, respectively. Also from these tables, one can observe the significant improvement of the proposed technique with respect to the state-of-the-art methods. The producer accuracies of the classes show that the most critical class is "Agriculture Fields", which is completely missed when training is done using only source domain samples, and its accuracy can only be partially recovered at the convergence of the proposed algorithm. The class "Forest" is also associated to very low accuracy at the first iteration, but its accuracy is significantly improved at the convergence.

TABLE III CLASSIFICATION RESULTS OBTAINED ON THE TEST SET TS_2 WITH THE PROPOSED DA TECHNIQUE AT DIFFERENT ITERATIONS (I.E., WITH DIFFERENT NUMBER OF TARGET-DOMAIN SAMPLES INCLUDED IN THE TRAINING SET). THE RESULTS ARE REPORTED IN TERMS OF PRODUCED ACCURACIES OF THE CLASSES, OVERALL ACCURACY AND KAPPA COEFFICIENT (MULTISPECTRAL VHR DATA SET).

	Number of target-domain samples included in the training set				
	0	50	100	150	200
Vineyard	74.7%	76.9%	73.1%	75.6%	83.1%
Water	99.8%	100.0%	100.0%	100.0%	100.0%
Agriculture Fields	0.0%	22.6%	70.2%	56.1%	34.8%
Forest	5.5%	73.9%	87.2%	88.0%	84.7%
Apple Tree	90.6%	96.4%	97.1%	97.0%	98.5%
Urban Area	100.0%	98.7%	91.5%	88.0%	94.6%
Overall Accuracy	69.2%	82.5%	84.1%	84.3%	87.1%
Kappa Coeff.	0.591	0.771	0.794	0.795	0.829

TABLE IV CLASSIFICATION RESULTS OBTAINED ON THE TEST SET TS_2 WITH THE WITH THE TECHNIQUE BASED ON THE COMBINATION OF $q+$ AND THE WEIGHING STRATEGY AT DIFFERENT ITERATIONS (I.E., WITH DIFFERENT NUMBER OF TARGET-DOMAIN SAMPLES INCLUDED IN THE TRAINING SET). THE RESULTS ARE REPORTED IN TERMS OF PRODUCED ACCURACIES OF THE CLASSES, OVERALL ACCURACY AND KAPPA COEFFICIENT (MULTISPECTRAL VHR DATA SET).

	Number of target-domain samples included in the training set				
	0	50	100	150	200
Vineyard	74.7%	75.8%	76.4%	76.5%	76.6%
Water	99.8%	100%	100%	100%	100%
Agriculture Fields	0.0%	15.2%	15.5%	16.2%	18.1%
Forest	5.5%	47.5%	65.9%	76.1%	76.1%
Apple Tree	90.6%	96.3%	96.6%	96.8%	96.7%
Urban Area	100%	100%	100%	100%	100%
Overall Accuracy	69.2%	77.8%	81.0%	82.6%	82.7%
Kappa Coeff.	0.591	0.708	0.750	0.772	0.774

Additional results obtained by changing the size of the initial training sets are reported in fig. 7 and 8. In accordance to our expectation, the best accuracies were obtained by using relatively small values of h^-, when starting from a small initial training set, and by using relatively higher values of h^- when the initial training set is large. In our experiments, the best results were obtained with $h^- = 4$ for the case of an initial training set size equal to 485; and with $h^- = 28$ for the case where its size is 1985. In the case where the initial training sets are smaller (fig. 7) we observe that state-of-the-art techniques converge to higher accuracies, as they are affected by a smaller number of source-domain samples. Conversely, in the case where the size of initial training sets is bigger (fig. 8) we observe that the standard techniques reach poorer classification accuracies, and the gap with respect to the proposed technique is larger.

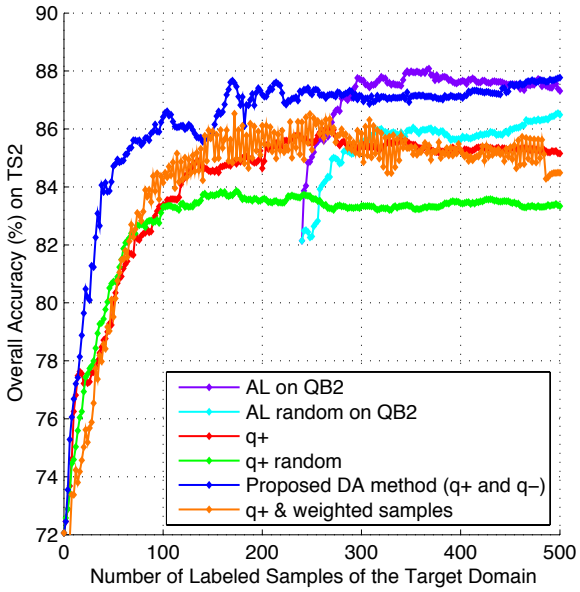


Fig. 7 – Overall classification accuracy (%) (averaged over ten trials) obtained on the test set TS_2 of the image QB_2 (target domain) versus the number of labeled samples of the target domain (selected from U) added to the training set by the different considered methods (multispectral VHR data set, initial training set size = 485).

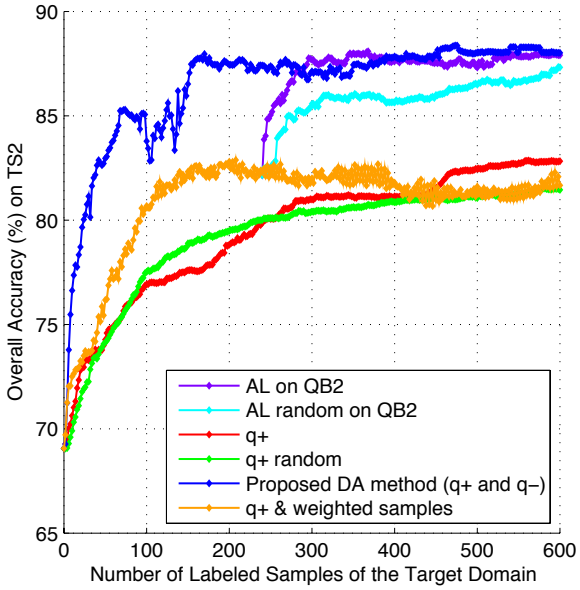


Fig. 8 – Overall classification accuracy (%) (averaged over ten trials) obtained on the test set TS_2 of the image QB_2 (target domain) versus the number of labeled samples of the target domain (selected from U) added to the training set by the different considered methods (multispectral VHR data set, initial training set size = 1985).

B) Results on the Hyperspectral Data Set

Fig. 9 displays the learning curves obtained for the hyperspectral data set. Like the previous case, the plot shows the averaged classification accuracies on TS_2 obtained with the different techniques versus the number of labeled samples of the target domain added to the training set. The reported curves are obtained fixing $h^+ = 10$ for all methods, and $h^- = 30$ for the proposed DA method. The optimum value of h^- was selected on the basis of several preliminary trials with $h^- = 10, 20, 30, 40, 50, 60$.

The results obtained on this second data set confirm the general trend observed in the previous one. The proposed DA technique resulted in higher accuracies on the test set TS_2 than standard AL methods using only the $q+$ function, random selection, or the $q+$ function combined with the sample-weighting mechanism. Also in this case, it is important to note that the proposed method reached the convergence with a significantly smaller number of samples than those required by applying AL directly to the target image. Indeed, a random selection of samples from target domain led to very unstable results up to 480 sample in our ten trials (this is clearly observable from the standard deviation evaluated over the ten trials), while the proposed approach converged to a stable average accuracy of 95% with only 230 samples. The proposed approach allows the user to find the best tradeoff between the classification accuracy and the number of labeled samples that can be acquired from target domain, which is related to the available budget for reference data collection. If the budget is limited, the user might decide to stop the AL procedure at early iterations. In our experiments, for example, the procedure could be stopped with just 120 samples, obtaining an accuracy of 91%. This result could not be achieved without considering the information available from the source domain, as the minimum number of samples for obtaining reasonable classification results on TS_2 by using only training samples on the Area 2 is 300.

Fig. 10 shows the average Bhattacharyya distance $B(i)$ between the class distributions computed using the initial training set $T^{(i)}$ and the training set $T^{(0)}$ (obtained using the proposed DA method) versus the number of labeled samples of the target domain included in $T^{(i)}$ (the Bhattacharyya distances $B_n(i)$ for the all considered classes $\omega_n \in \Omega$ are not reported for this data set because of the high number C of classes). Similarly to the VHR data set, considering the behavior of $B(i)$ we can observe that the proposed DA technique reached the converge when about 220 labeled samples of the target domain are included in the training set, in agreement with the learning curves depicted in Fig. 9. After 290 samples of the target domain have been included in the training set the distance $B(i)$ starts to slightly decrease. Using (10) to detect the saturation point of $B(i)$ (with s = 4 and $\varepsilon = 5e-3$), the algorithm is stopped

when 290 samples from target domain are added to the training set. At this point the overall accuracy is 95.5%, i.e., just 1% smaller than the maximum accuracy of 96.5% that is reached considering 390 samples from the target domain. Thus, we can conclude that also for this data set a reasonable stop criterion can be defined on the basis of the labeled samples of $T^{(0)}$ and $T^{(i)}$, without considering an independent test set defined on the target domain.

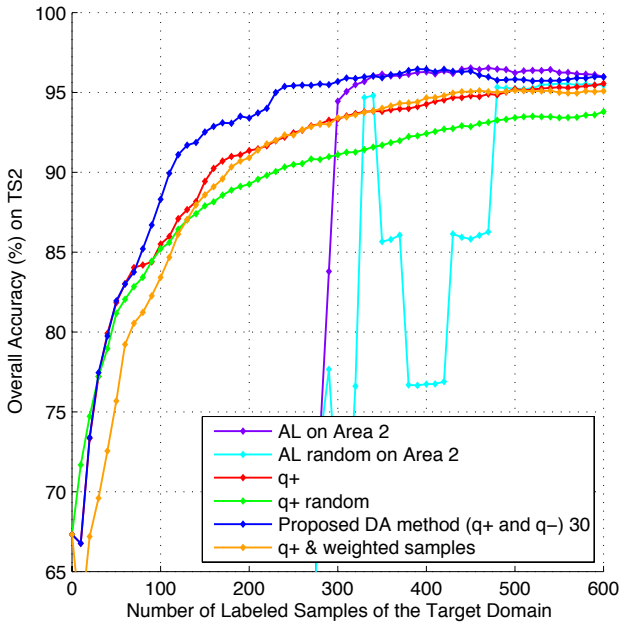


Fig. 9 – Overall classification accuracy (%) obtained on the test set TS_2 of the Area 2 (target domain) versus the number of labeled samples of the target domain (selected from U) added to the training set by the different considered methods. a) Averaged learning curves over ten trials. b) Averaged learning curves with standard deviation around the mean value (hyperspectral data set).

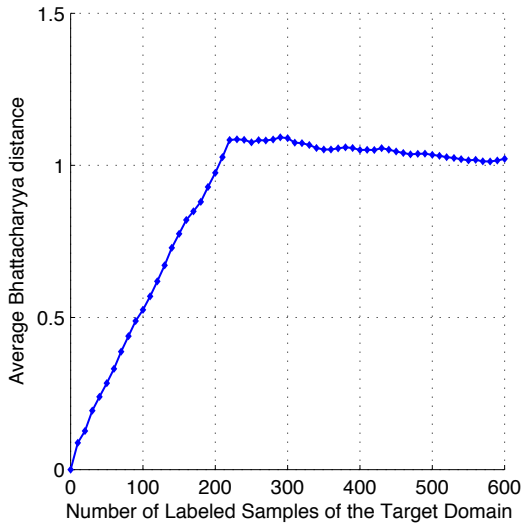


Fig. 10 - Average Bhattacharyya distance $B(i)$ between the class distributions computed on the initial training set $T^{(0)}$ and on the training set $T^{(i)}$ (obtained using the proposed DA method) versus the number of labeled samples of the target domain considered in $T^{(i)}$ (hyperspectral data set).

TABLE V and TABLE VI report the classification results at different iterations in terms of overall accuracy, kappa coefficient and producer accuracies of the considered classes obtained with the proposed DA technique and with the method based on the combination of $q+$ and the weighing strategy, respectively. In this data set, the classes associated with the highest shift between the two domains are "Acacia grasslands" and "Mixed mopane", for which the training with only source-domain samples led to an accuracy slightly higher than 20%. However, both classes reached very high classification accuracy at the convergence of the DA algorithm.

TABLE V CLASSIFICATION RESULTS OBTAINED ON THE TEST SET TS_2 WITH THE PROPOSED DA METHOD AT DIFFERENT ITERATIONS (I.E., WITH DIFFERENT NUMBER OF TARGET-DOMAIN SAMPLES INCLUDED IN THE TRAINING SET). THE RESULTS ARE REPORTED IN TERMS OF PRODUCED ACCURACIES OF THE CLASSES, OVERALL ACCURACY AND KAPPA COEFFICIENT (HYPERSPECTRAL VHR DATA SET).

	Number of target-domain samples included in the training set				
	0	50	100	200	300
Water	94.9%	96.7%	99.8%	100.0%	100.0%
Hippo grass	46.1%	100.0%	100.0%	100.0%	100.0%
Floodplain grasses1	46.7%	51.9%	87.1%	97.3%	97.5%
Floodplain grasses2	89.6%	93.5%	96.5%	99.6%	100.0%
Reeds1	79.2%	80.0%	79.4%	73.8%	83.2%
Riparian	79.8%	80.8%	84.2%	77.9%	84.4%
Firescar2	97.7%	97.7%	97.7%	99.1%	99.5%
Island interior	60.0%	100.0%	100.0%	99.7%	100.0%
Acacia woodlands	66.2%	71.5%	81.3%	92.0%	96.6%
Acacia shrublands	93.9%	87.4%	92.6%	93.7%	95.2%
Acacia grasslands	22.1%	99.5%	98.4%	97.7%	99.8%
Short mopane	98.9%	99.6%	95.6%	91.5%	93.0%
Mixed mopane	21.8%	39.2%	55.4%	93.7%	94.8%
Exposed soil	92.1%	100.0%	100.0%	100.0%	100.0%
Overall Accuracy	67.3%	81.9%	88.3%	93.4%	95.7%
Kappa Coeff.	0.647	0.805	0.873	0.928	0.953

TABLE VI CLASSIFICATION RESULTS OBTAINED ON THE TEST SET TS_2 WITH THE PROPOSED DA METHOD AT DIFFERENT ITERATIONS (I.E., WITH DIFFERENT NUMBER OF TARGET-DOMAIN SAMPLES INCLUDED IN THE TRAINING SET). THE RESULTS ARE REPORTED IN TERMS OF PRODUCED ACCURACIES OF THE CLASSES, OVERALL ACCURACY AND KAPPA COEFFICIENT (HYPERSPECTRAL VHR DATA SET).

	Number of target-domain samples included in the training set				
	0	50	100	200	300
Water	94.9%	83.7%	96.0%	98.2%	99.1%
Hippo grass	46.1%	90.6%	100.0%	100.0%	100.0%
Floodplain grasses1	46.7%	56.9%	72.3%	85.6%	89.8%
Floodplain grasses2	89.6%	89.6%	97.0%	97.0%	97.8%
Reeds1	79.2%	83.6%	85.4%	85.6%	84.6%
Riparian	79.8%	84.8%	81.3%	85.8%	89.2%
Firescar2	97.7%	97.7%	97.7%	98.0%	98.0%
Island interior	60.0%	76.8%	100.0%	100.0%	100.0%
Acacia woodlands	66.2%	37.9%	70.8%	85.1%	92.1%
Acacia shrublands	93.9%	76.5%	79.8%	84.1%	90.0%
Acacia grasslands	22.1%	99.7%	99.5%	99.8%	100.0%
Short mopane	98.9%	89.6%	93.7%	92.6%	92.6%
Mixed mopane	21.8%	42.3%	39.2%	78.9%	85.7%
Exposed soil	92.1%	100.0%	100.0%	100.0%	100.0%
Overall Accuracy	67.3%	75.7%	83.4%	90.9%	93.4%
Kappa Coeff.	0.647	0.737	0.821	0.901	0.929

VII. CONCLUSION

In this paper, a novel technique has been presented for addressing DA problems in the classification of remote sensing images with AL. Assuming that a remote sensing image and the related reference labeled samples are available from a previous analysis, the proposed technique can be used to classify another image acquired on another geographical area with similar characteristics and the same land-cover classes. Moreover, the proposed technique can be used also for updating the land-cover map given a new image acquired on the same area at a different time. In both operational scenarios, the proposed method allows the user to effectively exploit the information already available from the first remote sensing image (source domain) to classify the new scene (target domain) with the same set of information classes and correlated class distributions. The proposed technique is based on the definition of two query functions: 1) $q+$, which is devoted to select the most informative samples from the target domain on the basis of their uncertainty, and 2) $q-$, whose aim is to remove from the current training set source-domain samples which are not representative of the target-domain problem. It is important to note that $q+$ is associated with a standard notion of AL, while the concept of $q-$ is one of the main novel contributions of this work. The experimental results obtained on both a VHR multispectral and a hyperspectral data set confirm the effectiveness of the proposed technique. In particular, we observed on the two considered data sets that the proposed DA technique led to significantly higher accuracies on the target domain compared to standard AL methods using either only the $q+$ function, a random selection, or a state-of-the-art DA technique based on a sample-weighting strategy [16]. Moreover, it reached the convergence with a significant smaller number of samples than those required by applying AL directly to the target image. The proposed approach allows the user to find the best tradeoff between the classification accuracy and the number of labeled samples that can be acquired from target domain, which is related to the available budget for reference data collection.

Regarding the free parameters of the proposed method, we observed that the value of h^+ should be set to a small value, given that no diversity criterion is considered in the proposed technique. The optimal performance can theoretically be obtained with $h^+ = 1$ (in order to avoid possible redundancy among the batch of samples selected at a given iteration), but to reduce the computational burden and the possible overhead due to repeating the labeling procedure, this value can be greater than one. On the basis of our experimental analysis, we suggest to set the value of h^+ in the range [1,10]. The value h^- should be theoretically set considering the similarity between source and target domains. In case the two

domains are similar, a small value of h^- is sufficient, because only few samples from the source domain should be removed from the training set. If $P^t(y|\boldsymbol{x}) \approx P^s(y|\boldsymbol{x})$ the value of h^- should tend to zero. Conversely, if the two domains are significantly different, and in particular if $P^t(y|\boldsymbol{x})$ is significantly different from $P^s(y|\boldsymbol{x})$, the value of h^- (and therefore that of α) should be large. In this case, the value of h^- depends also on the size of T_s: the higher is the size of T_s, the larger the value of h^- should be set. Operationally, we suggest to set it in the range such that $\alpha \in [1,15]$.

An important concept that we introduced in this paper is that it is possible define a stopping criterion for the DA technique based on the behavior of the Bhattacharyya distances between the class distributions computed on the initial training set and the training set at the *i*-th iteration, without the need for test set TS_2 defined on the target domain. This is a very important result because in real DA problems it is not realistic to have test samples for the target domain.

Regarding the use of the proposed technique in real applications, we believe the it can be effectively adopted for optimizing the collection of reference labeled data for the classification of a new image in both the cases where the labeling have to be carried out by 1) visual image photointerpretation, or 2) *in situ* ground survey. In the first case, the photointerpreter can be guided *on-line* by the proposed AL algorithm to select the most informative samples of the target domain to be labeled and added to the training set. In the second case, the interactive AL algorithm can still be used for an *on-line* guidance of the labeling process, without requiring the users to repeat different field campaigns (i.e., without requiring additional travel to the study area). This can be done considering a set-up where the user performs the field survey with the aid of a GPS device connected to a laptop or a mobile portable device that allows him to annotate the reference information and run the proposed AL algorithm that can guide him on which new sample to label. In case the processing is too demanding, the mobile device can be used to send the new acquired information on a mobile network to a remote station where the analysis of the data take place. The remote station can send prompt feedback (exploiting the proposed technique) to the mobile device of the user, guiding him on the next samples to label. This system allows the user to largely reduce the cost and time associated to the classification of the second image. As an important remark it is worth noting that the proposed AL method can also be integrated with more traditional sampling strategies for collecting training samples, by considering the output of the classifier for identifying the most important samples in a given area.

A limitation of the proposed technique is given by the fact that the adopted $q+$ function exploits only an uncertainty criterion, but it does not consider the diversity of the samples (leading to the

selection of possible redundant samples if $h^+ > 1$). Thus, as future developments of this work we plan to include a diversity criterion in the $q+$. Another possible development could be the introduction of an automatic strategy for setting an adaptive value of h^- at each iteration.

ACKNOWLEDGMENTS

This work has been supported from the Autonomous Province of Trento. The authors would like to thank Prof. M. Crawford (Purdue University, W. Lafayette, IN) for kindly providing the hyperspectral data set.

REFERENCES

[1] P. Mitra, B. U. Shankar, and S. K. Pal, "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern Recognition Letters*, Vol. 25, No. 9, pp. 1067-1074, Jul. 2004.

[2] 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.

[3] D. Tuia, F. Ratle, F. Pacifici, M. Kanevski, and W. J. Emery, "Active Learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, Vol. 47, No. 7, pp. 2218 -2232, Jul. 2009.

[4] B. Demir, C. Persello, L. Bruzzone, "Batch Mode Active Learning Methods for the Interactive Classification of Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, Vol. 49, No.3, pp. 1014-1031, 2011.

[5] S. Patra, L. Bruzzone, "A Fast Cluster-Based Active Learning Technique for Classification of Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, Vol. 49, No.5, pp. 1617-1626, 2011.

[6] E. Pasolli, F. Melgani, and Y. Bazi, "Support vector machine active learning through significance space construction," *IEEE Geosci. Remote Sens. Lett.,* Vol. 8, No. 3, pp. 431–435, May 2011.

[7] W. Di, M. M. Crawford, "Active Learning via Multi-View and Local Proximity Co-Regularization for Hyperspectral Image Classification," *IEEE J. Sel. Topics Signal Process*, Vol. 5, No. 3, pp. 618-628, June 2011.

[8] B. Demir, F. Bovolo, L. Bruzzone, "Detection of Land-Cover Transitions in Multitemporal Remote Sensing Images With Active-Learning-Based Compound Classification," *IEEE Trans. Geosci. Remote Sens, in press.*

[9] L. Bruzzone, M. Marconcini, "Domain Adaptation Problems: a DASVM Classification Technique and a Circular Validation Strategy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 32, No. 5, pp. 770-787, 2010.

[10] H. Daumé III and D. Marcu, "Domain Adaptation for Statistical Classifiers," *J. Artificial Intelligence Research*, Vol. 26, pp. 101-126, 2006.

[11] L. Bruzzone, D. Fernandez Prieto, "Unsupervised Retraining of a Maximum-Likelihood Classifier for the Analysis of Multitemporal Remote-Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, Vol. 39, No.2, pp. 456-460, 2001.

[12] L. Bruzzone and D. Fernandez Prieto, "A partially unsupervised cascade classifier for the analysis of multitemporal remote-sensing images," *Pattern Recognition Letters*, Vol. 23, No. 9, pp. 1063-1071, 2002.

[13] L. Bruzzone and R. Cossu, "A Multiple-Cascade-Classifier System for a Robust and Partially Unsupervised Updating of Land-Cover Maps," *IEEE Trans. Geosciences and Remote Sensing*, Vol. 40, No. 9, pp. 1984-1996, Sept. 2002.

[14] L. Bruzzone, R. Cossu, and G. Vernazza, "Combining Parametric and Non-Parametric Algorithms for a Partially Unsupervised Classification of Multitemporal Remote-Sensing Images," *Information Fusion*, Vol. 3, No. 4, pp. 289-297, 2002.

[15] L. Gómez-Chova, G. Camps-Valls, L. Bruzzone, J. Calpe-Maravilla, "Mean Map Kernel Methods for Semisupervised Cloud Classification", *IEEE Trans. Geosci. Remote Sens.*, Vol. 48, No. 1, pp. 207-220, 2010.

[16] G. Jun, J. Ghosh, "An Efficient Active Learning Algorithm with Knowledge Transfer for Hyperspectral Data Analysis," *IEEE International Geoscience and Remote Sensing Symposium*, *IGARSS 2008*, Vol. 1, pp. I-52-I-55, 2008.

[17] D. Tuia, E. Pasolli, W. J. Emery, "Using active learning to adapt remote sensing image classifiers," Remote Sensing of Environment, Vol. 115, pp. 2232–2242, 2011.

[18] O. Chapelle, B. Schölkopf, and A. Zien, "Semi-supervised Learning," MIT Press, 2006.

[19] S. J. Pan, and Q.Yang, "A survey on Transfer Learning," *IEEE Transactions on Knowledge Transfer and Data Engineering*, Vol. 22, No. 10, pp. 1345-1359, October 2010.

[20] B. Zadrozny, "Learning and Evaluating Classifiers under Sample Selection Bias," *Proc. 21st Int'l Conf. Machine Learning*, 2004.

[21] M. Dudik, R.E. Schapire, and J.S. Philips, "Correcting Sample Selection Bias in Maximum Entropy Density Estimation," *Advances in Neural Information Processing Systems 17*, 2005.

[22] J. Huang, A. Smola, A. Gretton, K.M. Borgwardt, and B. Schölkopf, "Correcting Sample Selection Bias by Unlabeled Data," *Advances in Neural Information Processing Systems 20*, 2007.

[23] H. Shimodaira, "Improving Predictive Inference under Covariate Shift by Weighting the Loglikelihood Function," *J. Statistical Planning and Inference*, Vol. 90, pp. 227-244, 2000.

[24] M. Sugiyama and K.R. Müller, "Input-Dependent Estimation of Generalization Error under Covariate Shift," *Statistics and Decisions*, Vol. 23, pp. 249-279, 2005.

[25] H. Daumé III, "Frustratingly Easy Domain Adapatation," *Proceedings of the 2010 Workshop on DA for Natural Language Processing*, ACL 2010, Uppsala, Sweden, pp. 53-59, 15 July 2010.

[26] M. Li and I. Sethi, "Confidence-Based active learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 8, pp. 1251-1261, 2006.

[27] S. Rajan, J. Ghosh, and M. Crawford, "Exploiting Class Hierarchies for Knowledge Transfer in Hyperspectral Data," *IEEE Trans. Geosci. Remote Sens.*, Vol. 44, No. 11, pp. 3408-3417, Nov. 2006.

[28] W. Kim, M. Crawford, "Adaptive classification for hyperspectral image data using manifold regularization kernel machines," *IEEE Trans. Geosci. Remote Sens.*, Vol. 48, No. 11, pp. 4110-4121, 2010.

[29] G. Jun, J. Ghosh, "Spatially Adaptive Classification of Land Cover With Remote Sensing Data," *IEEE Trans. Geosci. Remote Sens.*, Vol. 49, No. 7, pp. 2662-2673, 2011.

[30] J. Jiang and C. Zhai, "Instance Weighting for Domain Adaptation in NLP," *Proc. 45th Ann. Meeting of the Assoc. for Computational Linguistics*, 2007.

[31] W. Dai, Q. Yang, G. Xue, and Y. Yu, "Boosting for transfer learning*," in Proceedings of the 24th International Conference on Machine Learning*, Corvalis, Oregon, 2007, pp. 193–200, ACM.

[32] Y. S. Chan and H. T. Ng, "Domain Adaptation with Active Learning for Word Sense Disambiguation," *Computational Linguistics,* Vol. 45, pp. 49-56, 2007.

[33] P. Rai, A. Saha, H. Daumé III, and S. Venkatasubramanian, "Domain Adaptation meets Active Learning," *Proceedings of the NAACL HLT 2011 Workshop on Active Learning for Natural Language Processing*, Los Angeles, California, USA, pp. 27-32, June 2010.

[34] L. Bruzzone, C. Persello, "A Novel Approach to the Selection of Spatially Invariant Features for the Classification of Hyperspectral Images with Improved Generalization Capability," *IEEE Trans. Geosci. Remote Sens.*, Vol. 47, No. 9, pp. 3180 – 3191, 2009.

[35] J. P. Hoffbeck and D. A. Landgrebe, "Covariance Matrix Estimation and Classification With Limited Training Data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 7, pp. 763-767, July 1996.

[36] M. Dalponte, L. Bruzzone, and D. Gianelle, "Fusion of hyperspectral and LIDAR remote sensing data for the classification of complex forest areas," *IEEE Transactions on Geoscience and Remote Sensing, Vol. 46, No. 5, pp. 1416-1427, May 2008.*

[37] 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. Machine Learning Research*, Vol. 6, pp. 589-613, 2005.

[38] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed. New York: Academic, 1990.