SVM-based Boosting of Active Learning Strategies for Efficient Domain Adaptation

Giona Matasci, *Student Member, IEEE*, Devis Tuia, *Member, IEEE*, Mikhail Kanevski

***Abstract*—We propose a procedure that efficiently adapts a classifier trained on a source image to a target image with similar spectral properties. The adaptation is carried out by adding new relevant training samples with active queries in the target domain following a strategy specifically designed for the case where class distributions have shifted between the two acquisitions. In fact, the procedure consists of two nested algorithms. An active selection of the pixels to be labeled is performed on a set of candidates of the target image in order to select the most informative pixels. Along the inclusion of the pixels to the training set, the weights associated with these samples are iteratively updated using different criteria, depending on their origin (source or target image). We study this adaptation framework in combination with a SVM classifier accepting instance weights. Experiments on two VHR QuickBird images and on a hyperspectral AVIRIS image prove the validity of the proposed adaptive approach with respect to existing techniques not involving any adjustments to the target domain.**

***Index Terms*—image classification, active learning, domain adaptation, TrAdaBoost, instance weights, SVM.**

I. Introduction

When dealing with supervised image classification, the collection of ground truth data is among the key factors influencing the quality of a land cover map. To be effective, a classifier needs examples suitably representing the spectral signature of the various classes found in the scene. Nevertheless, the sampling process is not a trivial task. The procedure requires either expensive terrain campaigns (usually when dealing with hyperspectral data of low spatial resolution) or time-consuming photo-interpretation analyses (utilized in particular when coping with very high resolution (VHR) images).

In such a context, *active learning* (AL) techniques have been widely studied in the remote sensing community during the last years [1]–[4]. Indeed, procedures allowing the user to optimally select the pixels to label can dramatically reduce the sampling burden. Smartly built training sets yield classification models efficiently discriminating the land cover classes. AL methods perfect the passive acquisition of labeled samples by providing the user with hints about the most informative pixels in the image. By assigning a label to these uncertain pixels, the classifier benefits of information where it most needs it for its direct improvement.

Manuscript received XXX 20XX; revised XXX 20XX; accepted XXX 20XX. This work has been supported by the Swiss National Science Foundation with grants no. 200021-126505 and PZ00P2-136827.

GM and MK are with Institute of Geomatics and Analysis of Risk, University of Lausanne, Switzerland. E-mail: {giona.matasci, mikhail.kanevski}@unil.ch,

DT was with the Image Processing Laboratory (IPL). Universitat de València, Spain. He is now with the Laboratory of Geographic Information Systems (LASIG), Ecole Polytechnique Fédérale de Lausanne, Switzerland. E-mail: devis.tuia@epfl.ch

Besides AL strategies, it has been shown that the labeling effort could be further reduced by re-utilizing already collected ground truth associated with images acquired by the same sensor in a region with comparable characteristics. This is done through the adaptation of a classifier designed to model a given image, the *source domain*, in order to model the image of interest, the *target domain*. The target image shares the same spectral channels and classes to be described but its pixels are assumed to be drawn from slightly different but related probability distributions: a *dataset shift* is said to have occurred [5].

The study of adaptation algorithms is referred to as *domain adaptation* (DA). This research field falls under the broader field of *transfer learning* [6]. For its part, the remote sensing community has quickly seen a growing interest in such procedures allowing to map and update the land cover over vast geographical areas by reusing existing models and ground reference data [7]–[14]. In [7], pixels of the target domain are used to re-estimate parameters of the maximum likelihood classifier. This way, the Gaussian clusters are matched to the data observed in the target domain. In [8], ensembles of classifiers are used to adapt to the target domain. Diversity in the predictions of the ensemble is used to reduce the number of trees. Bruzzone and Marconcini, in [9], propose to deform a SVM classifier by discarding contradictory old training samples with respect to the distribution observed in the target domain. At the same time, semi-labeled target samples are added to the training set. In [10], knowledge transfer is performed by matching the means of data clusters in a kernel-induced space. The authors of [11] use manifold regularization to adapt the model: the proposed algorithm forces the classification boundary to stay close to the low density region between two clusters of the target space. Jun and Ghosh, in [12], use spatial detrending with a Gaussian Process regression to compensate for spectral shifts that may have occurred in distinct regions of the image. In [13], an adaptation procedure aimed at finding a correspondence between the data manifolds via graph matching is introduced. Finally, in [14], the authors investigate the feature extraction framework in order to reduce the divergence between pixel distributions in the two domains. In both these last two cases the goal is to allow the direct application of a source classifier in the transformed target domain.

All the strategies reviewed above, however, assume that the labeled examples from the target image, when available, are passively obtained at once. On the contrary, if little resources can be allocated to the sampling and labeling of a given amount of new pixels, such sampling must be handled with care, in order to get maximal information from the limited

achievable queries. In this sense, the combined use of AL and DA approaches can be a winning strategy, since AL can be used to sample where the target image has shifted.

Recent advances in machine learning show that the combination of these two frameworks is effective: in [15] the authors proposed a scheme to label the most uncertain samples in the target domain for video classification. In [16], a principle apt to reduce the number of examples to be labeled by the user is outlined. The authors suggest to use the knowledge transferred from the source domain to label relevant target instances highlighted using AL. Later on, Rai *et al.* proposed a preprocessing step highlighting the interesting regions of the target domain in order to reduce the size of the set of candidate samples [17]. Based on this contribution, the same authors provided a complete framework for AL in a DA setting [18].

The interest of these approaches in remote sensing data classification is straightforward: a classifier trained on a first acquisition can be adapted to a new image with minimal effort by finding the pixels representing the shift between the two images. Such a principle was firstly explored by Jun and Ghosh [19]. The authors showed that DA could be achieved by actively querying, pixel by pixel and using a method constrained by data normality assumptions, the target samples necessary to be integrated in the knowledge transfer process. In [20], Tuia *et al.* proposed AL for the correction of sample selection bias when dealing with large images and unknown classes in the target domain. Successively, other methods specifically designed to reuse already collected ground truth information to initialize the AL loop have been advised [21].

In this paper, we propose to efficiently combine the DA and AL frameworks in the context of *Support Vector Machine* (SVM) classification [22], a supervised learner widely investigated in the recent years by the remote sensing community [23], [24]. The most informative pixels are sampled with active queries from the target image while adapting the obtained classifier using a transfer learning strategy, *TrAdaBoost* [25], to leverage the original source data. This principle, taken as starting point in [19], promotes a re-weighting of the training instances provided to the classifier in order to attribute a broader impact to key target domain samples while decreasing the influence of misleading source samples. This last step boosts the performance of traditional AL techniques when asked to intelligently suggest a sampling scheme in a target image whose class distributions have shifted.

The present contribution provides a thorough illustration of the TrAdaBoost algorithm and an analysis of its behavior. The adaption via this boosting technique is explored when the latter is run in combination with a classifier not requiring any assumption for the class-conditional statistical distributions. Indeed, a version of the SVM classifier integrating weights for the instances is analyzed by exploring the effectiveness of its adjustments aimed at meaningfully handling the new distribution of the data. We study the separate evolution, with respect to the domain of membership, of the support vectors and their weights during the AL procedure.

Additionally, we carried out experiments studying the individual impact of the two approaches combined here: active queries and samples re-weighting. From the results, we can appreciate how both approaches are complementary and perform differently depending on the degree and complexity of the shift. Still, in all experiments, their combination resulted in an improved solution always providing the best results.

The sampling strategies are tested on two datasets. The first one concerns two QuickBird images of urban scenes while the second one implies a hyperspectral AVIRIS image of a natural environment. In both cases, experimental results prove the efficacy of the technique with respect to traditional non-adaptive AL approaches.

The proposed methodology is outlined in Section II. Section III describes the datasets used and the setup of the experiments, while Section IV reports and discusses the results. Finally, Section V summarizes the main achievements of this work.

II. Adaptive Active Learning

This section presents the proposed algorithm, that combines AL and DA to compensate for a shift occurred between two image acquisitions. We refer to this framework as *adaptive active learning*. The purpose is to build a classifier that is able to efficiently handle the samples coming from the new image in order to provide a more accurate and adapted AL criterion. Such a criterion, inducing the labeling of more informative target pixels, is thus intended to boost the classification performance in the target domain.

As base AL heuristic we apply the *breaking ties* strategy (BT) [26]. BT uses posterior class probabilities to rank the potential new training samples according to their uncertainty for the current model. In this contribution SVM posterior probabilities estimated with the Platt's method [27] are considered. We then combine this AL strategy with a transfer learning technique known as TrAdaBoost, initially presented in [25]. Such a method achieves the desired adaptation via instance re-weighting: this paper proposes an analysis of its performance when combined with a SVM classifier accepting, in the optimization phase, weights associated with the training data samples. Note that, however, the choice of the base classifier is not restricted to SVM. Other classifiers allowing sample weights (e.g. LDA) can be used. Similarly, the AL procedure can be run using the sample selection heuristic that best suits the needs of the user.

A. SVM using Instance Weights

In this section, the base classifier utilized in this AL study is introduced. As it will be explained next, the proposed boosting technique requires the training examples to be weighted by the classifier. Working with SVM, theoretical descriptions of the implementation accepting these weights are presented in [28] for classification purposes as well as in [29] for regression tasks. The main points differentiating it from the standard version are detailed hereafter.

In the weighted variant of the SVM, during the optimization, one assigns sample weights $\mathbf{w} = \{w_i\}_{i=1}^{n}$, $w_i \in \mathbb{R}^+$ to all the n training samples belonging to the training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$.

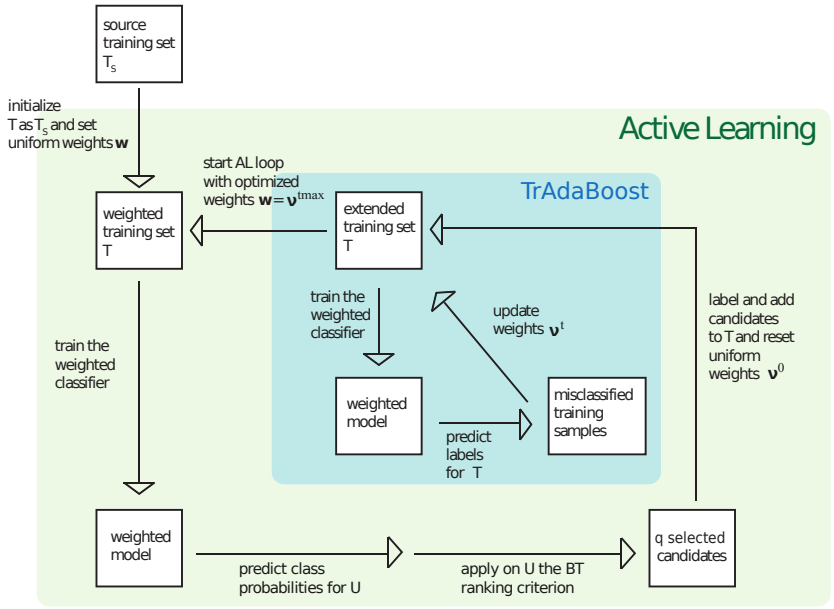


Fig. 1. Adaptive AL scheme: the outer AL loop is highlighted in green, while the inner TrAdaBoost loop is highlighted with blue tones.

Then, the training of the weighted SVM implies the solving of the following primal problem

$$\min_{\mathbf{v},b,\boldsymbol{\xi}} \left\{ \frac{1}{2}\|\mathbf{v}\|^2 + C\sum_{i=1}^{n} w_i \xi_i \right\} \tag{1}$$

subject to

$$y_i(\mathbf{v}^\top \mathbf{x}_i + b) \geq 1 - \xi_i\,, \quad i = 1,\ldots,n\,, \tag{2}$$

$$\xi_i \geq 0\,, \quad i = 1,\ldots,n\,, \tag{3}$$

where $\mathbf{v}$ is the vector defining the separating hyperplane, b is the associated bias term, ξ_i are the magnitudes of the permitted training errors and C is the usual penalty parameter determining the trade-off between margin maximization and training error minimization.

The associated dual problem is then set up as

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j \right\} \tag{4}$$

subject to

$$\sum_{i=1}^{n} y_i \alpha_i = 0\,, \tag{5}$$

$$0 \leq \alpha_i \leq w_i C\,, \quad i = 1,\ldots,n\,, \tag{6}$$

where the α_i's are the Lagrange multipliers related to each training point in the final (linear) SVM decision function

$$f(\mathbf{x}) = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i^\top \mathbf{x} + b\,. \tag{7}$$

One can notice the upper-bound for such coefficients defining the actual influence of the support vectors (training points with $\alpha_i > 0$) being dependent on the sample weight w_i. This induces an increased flexibility of the method, with samples allowed to receive α_i coefficients larger than the employed C value when $w_i > 1$. Consequently, particularly relevant samples could have an additional impact on the classification system if compared to the usual SVM implementation.

B. TrAdaBoost and Active Learning

To achieve DA through AL, two nested loops are run in order to i) select the most useful samples in the target image (outer AL loop) while ii) iteratively adapting the resulting classifier to the new domain (inner TrAdaBoost loop). The scheme of Fig. 1 outlines the general procedure while Algorithm 1 provides details about its main steps. In the following, the two phases of the algorithm are described and their objectives are highlighted.

Initially, the available labeled training set T is composed of n source samples only, *i.e.* $T = T_S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$. We provide the active learner with a set of unlabeled target domain candidates $U = \{\mathbf{x}_j\}_{j=1}^{l}$ among which to choose interesting samples to be labeled by the user. These examples progressively extend the set of the m target training instances $T_T = \{(\mathbf{x}_i, y_i)\}_{i=n+1}^{n+m}$ initialized as $T_T = \{\}$.

Moreover, we start sample weights as $w_i = 1, \forall i$. We employ this initialization instead of that with uniform weights $w_i = \frac{1}{n}, \forall i$ [25], to let the second term of (1) become $C\sum_{i=1}^{n} \xi_i$, as in the usual SVM formulation.

- The outer loop of the adaptation procedure is an AL routine where, at each iteration, the q most interesting candidates $\mathbf{x}_j \in U$ are identified using the BT strategy and, after the assignment of the corresponding true label y_j, added to T_T. This heuristic selects the best points $\hat{\mathbf{x}}^{BT}$ according to the following ranking criterion [26]:

$$\hat{\mathbf{x}}^{BT} = \arg\min_{\mathbf{x}_j \in U} \big(\max_{cl \in \Omega} p(y_j^* = cl|\mathbf{x}_j) - \max_{cl \in \Omega \setminus cl^+} p(y_j^* = cl|\mathbf{x}_j) \big), \tag{8}$$

Algorithm 1 Adaptive AL with TrAdaBoost

Inputs: initial source training set $T_S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, set of target domain candidates $U = \{\mathbf{x}_j\}_{j=1}^{l}$, number of candidates to add at each iteration q, number of TrAdaBoost iterations *tmax*
initialize $T_T = \{\}$, *i.e.* $m = 0$
initialize $T = T_S$
initialize $\mathbf{w}$ with equal unit weights $w_i = 1, \forall\, i$
for each AL iteration **do**
    train the SVM using T (weighted by $\mathbf{w}$) as training set
    compute the SVM test accuracy in the target domain
    predict the c class probabilities $p(y_j^* = cl|\mathbf{x}_j)\, \forall\, \mathbf{x}_j \in U$
    compute ranking criterion according to Eq. (8)
    remove the best q candidates from U and add them to T_T
    set $T = T_S \cup T_T$
    set $\boldsymbol{\nu}^t = \boldsymbol{\nu}^0$ to unit weights $\nu_i^t = \nu_i^0 = 1, \forall\, i$
    for each TrAdaBoost iteration $t = 1, \ldots, tmax$ **do**
        train the SVM (weighted by $\boldsymbol{\nu}^t$) using the extended T
        repredict the class labels $\hat{y}_i\, \forall\, \mathbf{x}_i \in T$
        calculate the weighted error ϵ_t on T_T according to Eq. (9)
        update weights to obtain $\boldsymbol{\nu}^{t+1}$ following Eq. (10)
    end for
    set $\mathbf{w} = \boldsymbol{\nu}^{tmax}$
end for
Outputs: final training set T, test classification accuracy along the AL iterations

where $cl^+ = \arg\max_{cl \in \Omega} \left(p(y_j^* = cl|\mathbf{x}_j)\right)$ is the class with the highest probability for pixel $\mathbf{x}_j$ and $\Omega = \{cl_1, \ldots, cl_c\}$ is the set of c classes. These probabilities are the output of the SVM classifier weighting the samples by means of vector $\mathbf{w} = \{w_i\}_{i=1}^{n+m}$. After the inclusion of the best candidate points to T_T, the complete training set T is updated as $T = T_S \cup T_T$.

- At each AL iteration, the inner TrAdaBoost loop is run to reweight the training instances in T. After adding the new labeled training samples, we initialize a new weighting vector $\boldsymbol{\nu}^t$ by setting equal weights $\nu_i^0 = 1, \forall\, i$ at the boosting iteration $t = 0$. Then, for every round of the inner loop, the labels $\hat{y}_i$ predicted by the current SVM model for the training samples are considered. In the multi-class case (extension of the binary problem approached in [25]), the weighted training error on the target set T_T is then computed as:

$$\epsilon_t = \sum_{i=n+1}^{n+m} \frac{\nu_i^t \cdot e_i}{\sum_{i=n+1}^{n+m} \nu_i^t} \tag{9}$$

where e_i takes a value of 1 if the classifier commits an error ($\hat{y}_i \neq y_i$) when labeling $\mathbf{x}_i$ and 0 otherwise ($\hat{y}_i = y_i$). Afterwards, the weights ν_i^t are updated for the subsequent boosting iteration in two distinct ways according to the domain of origin of $\mathbf{x}_i$, as proposed in [25]. In fact, we apply

$$\nu_i^{t+1} = \begin{cases} \nu_i^t \beta^{e_i} & \text{if } \mathbf{x}_i \in T_S \\ \nu_i^t \beta_t^{-e_i} & \text{if } \mathbf{x}_i \in T_T\,, \end{cases} \tag{10}$$

where

$$\beta = 1/(1 + \sqrt{2 \ln n / tmax})\,, \tag{11}$$

$$\beta_t = \epsilon_t/(1 - \epsilon_t)\,. \tag{12}$$

The process is run for *tmax* iterations and the final weights $\boldsymbol{\nu}^{tmax}$ are used to retrain the SVM with instance weighting ($\mathbf{w} = \boldsymbol{\nu}^{tmax}$), yielding the predictions in the target domain (test set and unlabeled candidates set). The associated estimated class probabilities are subsequently used by BT to perform the active selection on the pool of candidates U.

Taking a closer look at the TrAdaBoost loop, in Eq. (10), one will notice that if the sample is correctly classified the weight remains unchanged, whereas if the sample is misclassified, two options are possible. If the sample comes from the source domain, its weight is decreased by a constant factor (11). On the contrary, if the instance originates from the domain of interest, the target domain, its weight is increased by a factor inversely proportional to the target training error (12). This updating strategy aims at reducing the impact of misleading source examples, supposed to be the most dissimilar to the target instances the model should focus on. Conversely, the increase of the influence of misclassified target samples translates the need to concentrate on the regions of the target domain in which the class discrimination is harder. In light of these considerations, the boosting loop could be prone to overfit potential outliers. However, let us remark that, when the weighted target training error ϵ_t is excessively large (> 0.5), the reweighting factor β_t exceeds the value of 1, allowing therefore a decrease of the weights for the misclassified target samples in (10).

This transfer learning approach enables the SVM model to gradually adjust itself to the new domain. The different weighting of the examples leads to a boosted decision function more and more suited to model the input-output relationships in the target domain. Hence, the benefits of this procedure are twofold. On the one hand, the quality of the classification on test data (belonging to the target domain) is improved. On the other hand, since we are acquiring samples representing the target distribution, the class membership probabilities for the unlabeled samples in U are more accurately computed. This induces a selection criterion better suited to identify candidates lying in uncertain regions of the extended input space.

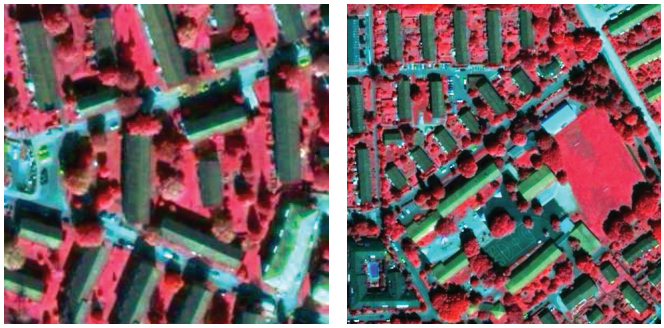
III. Data and Experimental Setup

In the following sections, the images considered, as well as the related setup of the experiments, are described. The proposed methodology has been tested on two datasets. The first one represents an urban case study bearing a moderate shift between the source and target images. On the contrary, in the second dataset the target domain is represented by a region showing remarkable differences in the spectral signatures of the vegetative cover with respect to the source region.

A. VHR QuickBird Images of Zurich

The first dataset consists of two VHR QuickBird images (acquired in 2002 and 2006) of the city of Zurich (Switzerland), representing two spatially distant neighborhoods. The target image was acquired in August while the source was acquired in October. The class-conditional distributions are affected by three factors: i) differences in illumination conditions, ii)

seasonal effects affecting vegetation growth and iii) varying materials composing roofs and roads. Fig. 2 illustrates the two considered images.



(a) (b)

Fig. 2. False color IR composites (RGB: bands 4-3-2) of the QuickBird images of the city of Zurich. (a) Source image (301×296 pixels). (b) Target image (474×482 pixels).

The pansharpened images have a spatial resolution of 0.6 m and present 4 bands covering the region of the spectrum from 450 to 900 nm. The histograms have first been matched and, subsequently, textural (3×3 data range, mean, homogeneity and entropy) and morphological (5×5 opening and closing, 7×7 and 9×9 opening and closing by reconstruction) features have been extracted from the panchromatic band to enrich the ground cover description with spatial information. The total number of considered features is 15 (4 MS bands, 1 PAN band, 4 textural, 6 morphological). Prior to the analyses, the variables have been normalized to have zero mean and unit variance, based on the source image descriptive statistics.

By visual inspection, we identified and labeled pixels from 5 classes characterizing both images: "Buildings", "Roads", "Grass", "Vegetation" and "Shadows". The training set for the source image is composed of 15'934 pixels while the unlabeled set of candidates extracted from the target image includes 22'723 pixels. The generalization ability of the different techniques in the target domain has been assessed on 26'797 test samples issued from spatially separated regions of the target image.

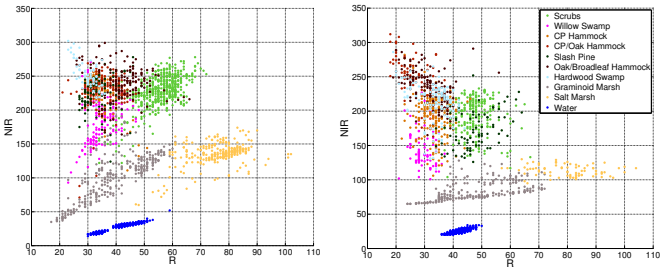
B. Hyperspectral AVIRIS Image of the KSC

The second case study is composed by two sub-regions of the same acquisition that has been obtained over the Kennedy Space Center (KSC), Florida (USA), on March 23, 1996 [1]. The images have been acquired with the AVIRIS hyperspectral instrument and are composed by 224 bands covering the region between 400 and 2500 nm. After the removal of water absorption and low SNR bands, the dataset was counting a total of 176 bands. The spatial resolution of the images is 18 m.

For the classification task we took into account only the land cover classes that are found in both images. The list of these classes, mainly consisting of types of subtropical vegetation, is given in Tab. I. As depicted by the scatterplots of Fig. 3, the classes present a rather large spectral variation across the two retained areas, justifying the definition of a source and a target domain. After a histogram matching procedure, the bands have been normalized (zero mean and unit variance) using source image parameters. A training set made up of 2'522 pixels was then issued from the source image. For the target image, we partitioned the available dataset into an unlabeled set of candidates and a test set both including 1'927 pixels. This target test set is then used for the comparison of the performances of the AL strategies.

TABLE I
CLASS NAMES AND SIZES (# OF LABELED PIXELS) FOR THE KSC DATASET.

Class name	Source image	Target image
Scrub	761	422
Willow swamp	243	180
CP hammock	256	431
CP/Oak hammock	252	132
Slash pine	161	166
Oak/broadleaf hammock	229	274
Hardwood swamp	105	248
Graminoid marsh	431	453
Salt marsh	419	156
Water	927	1392



(a) (b)

Fig. 3. Scatterplots for the two KSC images in the red (AVIRIS band # 29: $\approx$ 667 nm) VS near-IR (AVIRIS band # 49: $\approx$ 831 nm) space. The dataset shift observed from one image to the other is evident (divergence in both class-conditional and marginal probability distributions). (a) Source training set. (b) Target test set.

C. Experimental Setup

The experiments were conducted with 10 different and independent initializations of the training sets. For the Zurich images 1000 randomly selected pixels were retained, while the set size was fixed to 500 for the KSC dataset. A linear SVM has been used as supervised learner and a 5-fold cross-validation has been performed to find the optimal initial C parameter (extensive search in the space $\{0.1, \ldots, 100000\}$). For both datasets and for all the AL methods, $q = 10$ target samples per iteration were added to augment the initial source training set while the AL process was run for 35 iterations. At each iteration, the performance of the SVM models has been assessed on the test set extracted from the corresponding target image.

We compared the proposed adaptive active learning strategy (`AdaptiveAL_BT`) with the standard BT without instance reweighting (`AL_BT`) and with a procedure randomly selecting the pixels to label in the target image while adapting their weights following the TrAdaBoost scheme

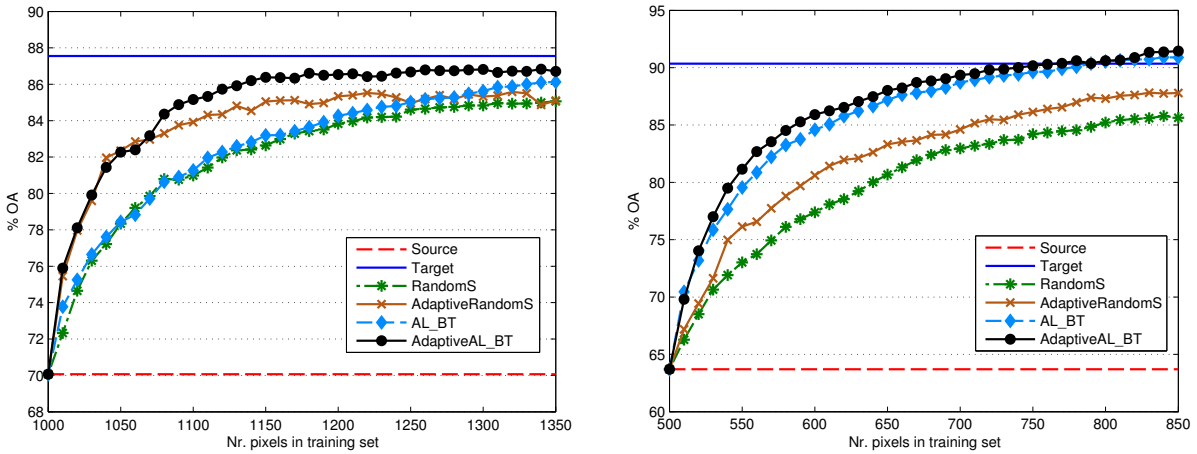


Fig. 4. Average learning curves (% OA) over 10 runs. `Source` (dashed red line) = model built using pixels of the source domain only, `Target` (solid blue line) = model built using pixels of the target domain only, `RandomS` (dashed green line with asterisks) = random sampling method, `AdaptiveRandomS` (solid brown line with crosses) = random sampling method combined with TrAdaBoost, `AL_BT` (dashed light blue line with diamonds) = AL via breaking ties, `AdaptiveAL_BT` (solid black line with circles) = proposed adaptive AL method. (a) Zurich target image. (b) KSC target image.

(`AdaptiveRandomS`). Also, in order to provide the usual AL baseline, the random selection of the samples to label (`RandomS`) has been considered. Finally, to set reference performances for the considered target images, linear SVM classifiers exclusively trained on target (pixels sampled from the set of candidates) and source (pixels sampled from the training set) datasets have been tested (`Target` and `Source` methods, respectively).

Regarding the proposed `AdaptiveAL_BT` method and `AdaptiveRandomS`, at each AL iteration, the weights of the samples in the training set were updated after 5 iterations of TrAdaBoost (stabilized ν_i values). In this sub-routine, the prediction on the training set was implemented through a 20-fold cross-validation to avoid overfitting.

The algorithms were implemented in MATLAB using LIBSVM [30] as library both for the standard SVM and instance weighting SVM (version available at `http://www.csie.ntu.edu.tw`). The computation of class probabilities to be used by BT is described in the same paper.

IV. Results

A. Learning Curves

Figure 4 summarizes the results for this task of DA through AL. The performance of the different AL techniques along the iterations (increasing training set size) has been assessed in terms of overall classification accuracy (OA). The depicted learning curves represent the average OA over the 10 experiments.

1) VHR QuickBird Images of Zurich: Analyzing Fig. 4(a), one can first notice the bad performance achieved by applying on the target data the source model (`Source`) without any adjustments (OA = 70.07%). The method consisting in randomly sampling the pool of unlabeled pixels (`RandomS`), considered as a baseline for AL, and the standard AL heuristic of BT both reveal a slow convergence. Nevertheless, the `AL_BT` method yields SVM models that are slightly more accurate than those built by sampling at random, but this happens only from the 10^{th} iteration onwards (approximately +0.5% OA).

The proposed combined methodology integrating the TrAdaBoost routine in the AL process (`AdaptiveAL_BT`) clearly outperforms these two sampling schemes by sharply increasing the classification accuracy since the very beginning of the AL iterations. In fact, already after 14 iterations (140 target pixels added) the associated curve achieves an OA of 86.2% (+3.4% with respect to `AL_BT`). Such a precision is never reached by the two baseline approaches during the considered first 35 AL cycles.

Nevertheless, we remark how the other procedure including the reweighting scheme, `AdaptiveRandomS`, is yielding a performance comparable to that of its active counterpart in the first 7 cycles of the AL routine. Subsequently, after the addition of 80 samples, the actively guided selection of the pixels to label provides an average improvement in OA of 1.5%.

It is interesting to note that none of the strategies is able to reach the `Target` performance at OA = 87.55%.

2) Hyperspectral AVIRIS Image of the KSC: Figure 4(b) reveals a similar pattern except for the improved performance of the `AL_BT` strategy and a worsen performance of the `AdaptiveRandomS` method.

The random sampling of the pixels in the target image (`RandomS`) results in poor updates of the initial training set. In fact, even after the inclusion of 350 samples, the model still lies 5% OA below the performance of a SVM trained with pixels from the target image only (`Target` reference classification with average OA = 90.35%).

On the other hand, both the `AdaptiveAL_BT` and the `AL_BT` show promising learning curves, eventually reaching and even exceeding the upper reference accuracy of the same-domain SVM model. In particular, one can remark the adaptive AL procedure evolving $\approx 1\%$ OA higher than its non-adaptive counterpart.

The effect of the intelligent selection of the most informative

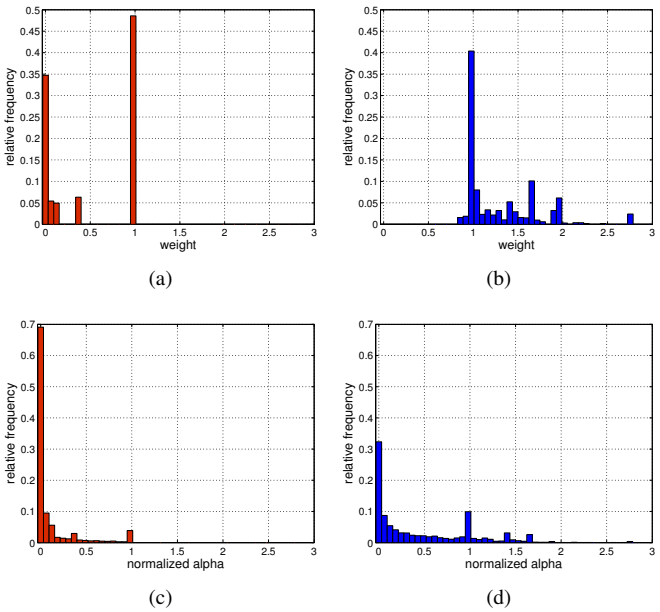


Fig. 5. Histograms showing the distribution of the final TrAdaBoost weights w_i and SVM coefficients α_i (normalized by C to account for its different values in the experiments) for the SVs of the two domains at the AL iteration #15 for the KSC dataset (frequencies over the 10 experiments). (a) Source SVs: weights w_i. (b) Target SVs: weights w_i. (c) Source SVs: α_i/C. (d) Target SVs: α_i/C.

pixels, as provided by the BT strategy, when combined with the TrAdaBoost algorithm is more evident on the KSC dataset. In fact, the curve associated with the integration of TrAdaBoost with the random, passive, sampling in the new image (`AdaptiveRandomS`) remains between 4 and 6% OA lower than the active one from the beginning of the AL process.

B. Analysis of Sample Weights

With the purpose to shed light on the actual effect of the TrAdaBoost model on the SVM-based AL procedure, it is worth analyzing the evolution of the weights w_i and the coefficients α_i along the AL iterations.

Figure 5 illustrates the distribution of the respective weights w_i and coefficients α_i for the support vectors (SVs) of each domain. Indeed, these are crucial training samples, the only ones contributing to the final SVM decision function. With this example concerning the KSC dataset, we focus on the state of the `AdaptiveAL_BT` method at the 15$^{\text{th}}$ AL iteration. From Figs. 5(a) and 5(b), one can observe how the weights of roughly 60% of the training SVs belonging to the source image are set to very low values ($w_i < 0.15$), whereas more than half of those of target domain SVs take values larger than 1. This translates, for the source SVs (Fig. 5(c)), to a significant amount of alpha coefficients found to be close to 0 and, for the target SVs (Fig. 5(d)), to a non-negligible number of alpha values that are actually larger than the corresponding SVM hyper-parameter C, *i.e.* $\alpha_i/C > 1$.

The highlighted pattern is noticeable since the early stages of the AL cycle, with more importance given to useful instances in the target domain and, conversely, with less weight assigned to misleading source instances. To better perceive the cited evolution as the AL and TrAdaBoost loops proceed to the adaptation of the SVM, we resort to Fig. 6.

Figure 6(a) depicts the evolution of the share of training points that eventually become SVs in the two domains. The number of such key samples remains stable over the entire AL procedure for the source training set T_S. On the contrary, for the target training set T_T we notice a growth of the considered ratio of SVs which is especially steep at first (until iteration 4), and then gradually slows down as the new image is sampled.

In Fig. 6(b), it is insightful to notice how, among the alpha coefficients (represented by their normalized counterparts α_i/C) associated with the SVs, there is a consistent polarization as the AL algorithm runs. In fact, always more and more of these α_i take either high values if corresponding to target samples, or low values if representing source samples. This evolution of the alphas is more marked for the target image, almost doubling the proportion of normalized $\alpha_i > 0.2$ found in the first iterations by the time the AL loop reaches its end. It is worth pointing out the sheer drop (from 65.2% to 39.7%) in the proportion of source normalized alphas larger than 0.02 when the first $q = 10$ target samples are added to the joint training set T.

C. Discussion

As pointed out in Sect. IV-A, an appropriately designed weighting scheme for the training instances, as the one provided by the presented method, ensures an optimal transfer of the knowledge between the source and the target image. A direct consequence of this fact, due to the improved posterior class probability estimates, is the more accurate selection of samples to be labeled along the AL iterations. We obtain an improved model, able to outperform in test the one built by selecting the training instances with the simple BT heuristic, if the latter is naively applied without any adaptation to the domain of interest. Moreover, improvements over the simple application of the TrAdaBoost algorithm in combination with a random sampling of the new image were observed. This highlights the impact of the active selection, via BT in this case, of the most helpful pixels of the target image.

In more detail, we can comment on the influence of the two qualities an adaptive AL system should possess: the ability to adapt to the domain of interest and the ability to actively select the new samples. The experiments we conducted reveal an opposite trend in the two considered datasets. On the one hand, on the Zurich images, we notice a higher importance given to the adaptation to the new domain (superior performance of the `AdaptiveRandomS` over the `AL_BT` method). That could be linked to the need of downweighting source pixels found in areas related to the shift, but in a rather stable environment, in terms of marginal distributions. At the same time, the misclassified target pixels, lying in a region where the class boundaries have changed, require more attention (increasing weights) to adjust the model. On the other hand, when dealing with the KSC dataset, the effect of the active sampling alone proves to be more decisive than the simple adaptation of the weights (superior performance of the `AL_BT` method). This behavior can be linked to the larger and nonlinear shift

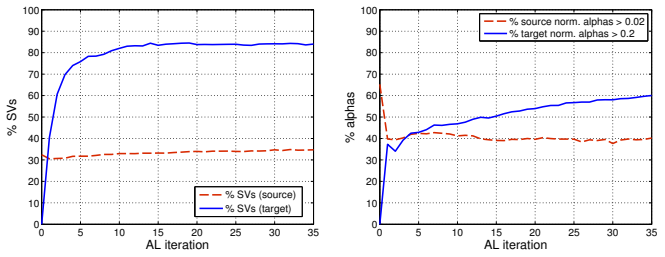
observed in this second hyperspectral case, that forces the algorithm to completely redefine the decision boundaries with the new queries. These additional samples are extremely useful to precisely redefine the new distribution of the highly mixed and overlapping classes that characterize the study area.

Despite the contradictory behavior observed in the case studies (in the first DA is more beneficial than AL, while in the second it is the contrary), the proposed method returns the best results in both cases. First, this illustrates the complementarity of the AL and DA approaches, that are effective in different scenarios. Second, this also strengthens the interest of a joint approach capable of taking the best from both worlds: in the nested loops of the proposed strategy, AL and DA interact constantly and can thus provide the relevant samples, while adapting the model to the new domain. The consequence is the remarkable gain in classification accuracy during the first iterations, observed in both case studies when using adaptive active sampling strategies.

Additionally, it is worth noting that, for the KSC images, both the active strategies converge to a performance exceeding that of the model built exclusively on target data. These superior classification accuracies are obtained with training sets that required the labeling of 270-290 target pixels only and thus showing the interest of intelligently built compact models avoiding the labeling of redundant samples. On the other hand, this fact indicates that the source data is still relevant and brings into play universal information that is useful to solve the problem in target. This accuracy improvement is even more significant in light of the large shift of the class spectral signatures existing between the two images, as testified by the $\approx 26.6\%$ OA difference between the `Source` and `Target` models (see also scatterplots of Fig. 3).

Section IV-B instead emphasizes the usefulness and the impact of the dedicated instance weights included in the SVM model, core of the proposed adaptive AL approach. As pointed out in Sect. II-A, the standard kernel learning machine optimizing the alpha coefficients with an equal upper-bound ($\alpha_i \leq C, \forall\, i$) is turned into an adaptable learning machine ($\alpha_i \leq w_i C, \forall\, i$). This weighted version of the SVM, as a matter of fact, is able to accord distinct relevance values to the training examples following both their domain of origin and their contribution to the class discrimination task. We draw the attention on the fact that, since these alpha coefficients act as sample weights in the SVM final decision function (7), the predictions are notably affected by the TrAdaBoost reweighting scheme.

In this sense, the evolution curves of Fig. 6 testify the increasing influence on the classification system of the pixels collected in the target image. As batches of these new domain samples are included in the training set, they quickly display a higher likelihood to become SVs than the already present source samples. Furthermore, the magnitude of the associated alpha coefficients is also increasing, translating the augmented relevance of the pixels belonging to the target domain we are interested in. Ultimately, the adjustable instance weights boost the SVM performance and enable the model to assign tailored alpha coefficients to its SVs. The AL process efficiently adapts the classifier by attributing more and more importance



(a) (b)

Fig. 6. Evolution along the AL iterations of the ratio of SVs and magnitude of the alphas for the two domains of the KSC dataset (percentages over the 10 experiments). (a) Percentage of SVs among source (dashed red line) and target (solid blue line) training data. (b) Percentage of source (dashed red line) and target (solid blue line) normalized alphas (α_i/C) larger than 0.02 or 0.2, respectively (percentages computed over the total number of alphas obtained in each domain).

to the target domain while discarding unprofitable source information. As a result, we obtain an improved discrimination of the land cover classes in the image for which we need to produce a new thematic map.

V. Conclusions

In this paper, an approach to boost the performance of active learning methods when applied in the context of domain adaptation has been presented and analyzed. We described a technique, TrAdaBoost, aimed at properly adapting training sample weights during the active learning process. Such adjustments proved potential in refining the ranking criterion for the selection of the most informative target pixels to be manually labeled by the user.

The individual contributions of the smart sampling and of the adaptive adjustment of sample weights have been assessed, concluding that the best performances are obtained when the two approaches are combined.

One of the objectives was also to uncover and better understand the behavior of the proposed reweighting scheme when integrated with a SVM classifier accepting instance weights in the training phase. The influence of these weights on the decision function, conveying the importance and pertinence to the domain of interest of each pixel, has been highlighted through the analysis of the evolution of the support vectors all along the sampling procedure. This way, useful insights have been provided concerning the significance of the SVM modification in order to integrate instance weights.

With the presented contribution, we demonstrated that, in a classification task involving a newly acquired image and when disposing of already collected ground truth data, the modeling effort for the target image can be efficiently reduced. In fact, by means of the proposed adaptive sampling strategy, the operator will be properly guided in the collection of the labels for the most useful pixels on the new image. As a consequence, standard supervised classifiers are supplied with a minimal and effective training set for a suitable land cover thematic mapping.

REFERENCES

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

[2] 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, 2009.

[3] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Muñoz-Marí, "A survey of active learning algorithms for remote sensing image classification," *IEEE J. Sel. Topics Sign. Proc.*, vol. 5, no. 3, pp. 606–617, 2011.

[4] B. Demir, C. Persello, and 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] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*, MIT Press, 2009.

[6] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[7] L. Bruzzone and D. F. 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.

[8] S. Rajan, J. Ghosh, and M. M. Crawford, "Exploiting class hierarchies for knowledge transfer in hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3408–3417, 2006.

[9] L. Bruzzone and M. Marconcini, "Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 4, pp. 1108–1122, 2009.

[10] L. Gomez-Chova, G. Camps-Valls, L. Bruzzone, and J. Calpe-Maravilla, "Mean map kernel methods for semisupervised cloud classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 1, pp. 207–220, 2010.

[11] W. Kim and 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.

[12] G. Jun and 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.

[13] D. Tuia, J. Muñoz-Marí, L. Gomez-Chova, and J. Malo, "Graph matching for adaptation in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. PP, no. 99, pp. 1–13, 2012.

[14] G. Matasci, M. Volpi, D. Tuia, and M. Kanevski, "Transfer component analysis for domain adaptation in image classification," in *Proc. SPIE Remote Sensing*, Prague, Czech Republic, 2011.

[15] J. Yang, R. Yan, and A. G. Hauptmann, "Adapting SVM classifiers to data with shifted distributions," in *Proc. IEEE Int. Conf. on Data Mining Workshops*, Washington, DC, USA, 2007, pp. 69–76.

[16] X. Shi, W. Fan, and J. Ren, "Actively transfer domain knowledge," in *Proc. Eur. Conf. Mach. Learn. - Principles Practice Knowl. Discov. Databases*, Berlin, Germany, 2008, pp. 342–357.

[17] P. Rai, A. Saha, H. Daumé III, and S. Venkatasubramanian, "Domain adaptation meets active learning," in *Proc. NAACL HLT Workshop on Act. Learn. Nat. Lang. Process.*, Los Angeles, CA, USA, 2010.

[18] A. Saha, P. Rai, H. Daumé III, S. Venkatasubramanian, and S. L. DuVall, "Active supervised domain adaptation," in *Proc. Eur. Conf. Mach. Learn. - Principles Practice Knowl. Discov. Databases*, Athens, Greece, 2011, pp. 97–112.

[19] G. Jun and J. Ghosh, "An efficient active learning algorithm with knowledge transfer for hyperspectral data analysis," in *Proc. IEEE IGARSS*, Boston, MA, USA, 2008, vol. 1, pp. I–52–I–55.

[20] D. Tuia, E. Pasolli, and W. J. Emery, "Using active learning to adapt remote sensing image classifiers," *Remote Sensing of Environment*, vol. 115, no. 9, pp. 2232–2242, 2011.

[21] C. Persello and L. Bruzzone, "Active learning for domain adaptation in the supervised classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. PP, no. 99, pp. 1 –16, 2012.

[22] V. Vapnik, *Statistical Learning Theory*, Wiley, 1998.

[23] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, 2004.

[24] 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, 2005.

[25] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proc. Intern. Conf. Mach. Learn.*, New York, NY, USA, 2007, pp. 193–200.

[26] T. Luo, K. Kramer, D. B. Goldgof, L. O. Hall, S. Samson, A. Remsen, and T. Hopkins, "Active learning to recognize multiple types of plankton," *J. Mach. Learn. Res.*, vol. 6, pp. 589–613, 2005.

[27] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*. 1999, pp. 61–74, MIT Press.

[28] G. H. Nguyen, S. L. Phung, and A. Bouzerdoum, "Efficient SVM training with reduced weighted samples," in *Proc. Intern. Joint Conf. Neural Networks*, San Jose, CA, USA, 2010, pp. 1–5.

[29] M. W. Chang, C. J. Lin, and R. C. Weng, "Analysis of switching dynamics with competing support vector machines," *IEEE Trans. Neural Networks*, vol. 15, pp. 720–727, 2004.

[30] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Giona Matasci (S'10) was born in Locarno, Switzerland, in 1985. He received the M.Sc. degree in environmental sciences from the University of Lausanne, Switzerland, in 2009. After an internship as research assistant at the LaSIG laboratory of the Federal Institute of Technology of Lausanne (EPFL), he is currently working toward the Ph.D. degree at the Institute of Geomatics and Analysis of Risk (IGAR) of the University of Lausanne. The topic of the thesis concerns machine learning and its applications to remote sensing and environmental data analysis with a focus on domain adaptation problems in image classification. Mr. Matasci serves as a reviewer for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS and IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.

Devis Tuia (S'07, M'09) was born in Mendrisio, Switzerland, in 1980. He received a diploma in Geography at the University of Lausanne (UNIL) in 2004, a Master of Advanced Studies in Environmental Engineering at the Federal Institute of Technology of Lausanne (EPFL) in 2005 and a Ph.D in environmental sciences at UNIL in 2009. He was a postdoc researcher at both the University of Valéncia, Spain and the Univeristy of Colorado at Boulder under a Swiss National Foundation program. He is now senior research associate at the LaSIG laboratory, EPFL. His research interests include the development of algorithms for information extraction and classification of very high resolution remote sensing images using machine learning algorithms. Visit `http://devis.tuia.googlepages.com/` for more information.

Mikhail Kanevski received the Ph.D. degree in plasma physics from the Moscow State University, Moscow, Russia, in 1984 and Doctoral thesis in computer science from the Institute of Nuclear Safety (IBRAE) of Russian Academy of Science, in 1996. Until 2000, he was a Professor at Moscow Physico-Technical Institute (Technical University) and head of laboratory at the Moscow Institute of Nuclear Safety, Russian Academy of Sciences. Since 2004, he is a professor at the Institute of Geomatics and Analysis of Risk (IGAR) of the University of Lausanne, Switzerland. He is a principal investigator of several national and international grants. His research interests include geostatistics for spatio-temporal data analysis, environmental modeling, computer science, numerical simulations and machine learning algorithms. Remote sensing and meteorological data analysis, natural hazards assessment (forest fires, avalanches, landslides) and time series prediction are the main applications considered at his laboratory.