*Manuscript

# Using active learning to adapt remote sensing image classifiers

D. Tuia[a,*], E. Pasolli[b], W. J. Emery[c]

[a]*Image Processing Laboratory (IPL), Universitat de València, Spain*
[b]*Information Engineering and Computer Science Dept., University of Trento, Italy*
[c]*Aerospace Engineering Dept., University of Colorado at Boulder, USA*

---

**Abstract**

The validity of training samples collected in field campaigns are crucial for the success of land use classification models. However, such samples often suffer from a sample selection bias and do not represent the variability of spectra that can be encountered in the entire image. Therefore, to maximize classification performance, one must perform adaptation of the first model to the new data distribution. In this paper, we propose to perform adaptation by sampling new training examples in unknown areas of the image. Our goal is to select these pixels in an intelligent fashion that minimizes their number and maximizes their information content. Two strategies based on uncertainty and clustering of the data space are considered to perform active selection. Experiments on urban and agricultural images show the great potential of the proposed strategy to perform model adaptation.

*Keywords:* Active learning, covariate shift, VHR, hyperspectral, remote sensing, image classification.

---

*Corresponding author. Tel.: +34 963544061; fax: +34 963543261
*Email address:* devis.tuia@uv.es (D. Tuia)

## 1. Introduction

Today, the access to remote sensing images has been made easier by the availability of images sensed by commercial satellites with short revisit periods. Sensors such as QuickBird or World-View II provide imagery at very high geometrical resolution, thus providing an unprecedented detail in the scenes described and allowing fine reconstruction of urban objects such as buildings. However, such a fine resolution leads to the increase of variability of the classes to be detected. For mid-resolution problem such as landuse classification, sub-metre resolution comes with strong intraclass variability caused by geometrical properties of the objects, changes in illumination and details detected only at the higher resolution (e.g., chimneys on buildings).

Even if they are able to treat well-defined classification tasks, the majority of current classification methods rely on supervision and may fail if the data used to build the model (the training set) are not representative of the true distribution generating the classes. Note that when dealing with remote sensing image classification, a user is often confronted with large archives of digital information to be classified and that the spatial extent of such images makes the definition of exhaustive training sets a difficult and time-consuming task. In this sense, providing exhaustive ground truth for large remote sensing images is often not possible. As a consequence, the labeled information only covers a part of the true variability of the class distribution. Moreover, a user can afford only partial ground surveys and can rely on previous studies about the ground cover. This is even more critical when adapting a model to a multitemporal sequence, where differences in illumination and reflectance can make the adaptation of a model fail (Rajan et al., 2006).

These constraints result in the user not having the economic and temporal resources to label the entire area or being confronted to a new classification task including a previously unconsidered and contiguous region in a second moment only. In both cases, one must then focus on subsections of the images, in order to retrieve a coherent training set representing the classes to be described and then apply the model obtained from the sub-image to the entire scene.

This field of investigation is primordial for remote sensing data analysis and has been considered for mid-resolution optical data as *signature extension*. In the pioneering paper by Fleming et al. (1975), the authors studied the effect of clustering the data to account for data multimodality in Gaussian classifiers, thus considering the issue of non-stationary data across the image. This principle has been applied in applications for Landsat imagery in Woodcock et al. (2001); Pax-Lenney et al. (2001); Foody et al. (2003); Olthof et al. (2005). In Jia & Richards (2002), the approach by Fleming et al. (1975) was successfully extended to hyperspectral data, thus showing the interest of considering model adaptation to unsampled areas for this type of imagery.

However, in recent methodological research this aspect has been overlooked by the focus put on the classification of local regions and by claiming that the new algorithms proposed were powerful enough to generalize to unseen areas. A common assumption in such developments became that data are homogeneous throughout the image, i.e. class statistics remain constant over the image. This seems unrealistic, especially when the training set only covers small subsets of the scene. In recent years, emphasis has been put on optimizing the classifiers for situations where the training set is minimal (Jackson & Landgrebe, 2001; Gómez-Chova et al., 2008; Camps-Valls & Bruzzone, 2009; Tuia & Camps-Valls, 2009), but the problem of adaptation to slightly varying test distributions has been considered only rarely in recent literature using spectral data. By this, we mean that a shift between the distribution of the training set and the test data has occurred, leading thus to an incompatibility of the model optimized for the first set of observations when they are used to describe the unseen pixels. In the machine learning community, the problem, also known as *covariate shift (Quiñonero-Candela et al., 2009),* has been considered from different perspectives: by weighting the observations according to the position of the training samples with respect to the support of the test ones (Sugiyama et al., 2007; Bickel et al., 2009) or by adding regularizers on the test data distribution (Yang et al., 2007). Covariate shift is being considered nowadays in several applications, covering brain computer interfaces (Li et al., 2010b) or genomic sequence analysis (Schweikert et al., 2008). In remote sensing literature, the field is relatively young: in Bruzzone & Fernandez-Prieto (2001), the samples in the new domain are used to assess the class parameters in the EM algorithm. In Rajan et al. (2006), a classifier built on an image is updated using the unlabeled data distribution of another scene in an hyperspectral image classification problem. In Bruzzone & Marconcini (2009), this idea is further developed with an iterative procedure adapting a training set to shifted images: the model discards contradictory old training samples and uses the distribution of the new image to adapt the model to the new conditions.

Finally, in Gómez-Chova et al. (2010), matching of the first order statistics in a projected space is studied under the name of kernel mean matching: the model is then applied to a series of images for cloud detection.

A strategy to learn the dataset shift is to sample additional pixels from the unknown distribution to check if they are consistent with the model obtained from training set generated by partial sampling. In particular, when dealing with very high resolution imagery, the problem of finding pixels lying in the shifted areas can be a difficult task. In this paper, we propose a simple, yet effective way to correct a training set for its application to a new area where a data set shift may have occurred. We propose to use active queries to learn the shift and sample the areas in which the classifier would become suboptimal, since they do not contain any labeled instance. These methods are new to the remote sensing community (Mitra et al., 2004; Rajan et al., 2008; Liu et al., 2008; Tuia et al., 2009b), but they are rapidly gaining interest in this community (see the recently published papers by Pasolli et al. (in press); Li et al. (2010a); Patra & Bruzzone (in press)), as they allow one to build an optimal training set with a minimum of queries (or labeled pixels).

Although appealing, the use of active learning for adapting a classifier to new data must be done carefully. Traditional supervised active learning algorithms focus on discrepancies near the classification boundary, resulting in new contradictory areas that may appear in the unseen distribution (the new image). However, such contradictions may happen far from those boundaries, for instance if a new class has appeared. In this case, an active learning algorithm risks failure and can lead to slower convergence than random sampling that may find these regions by chance.

In this paper, we study the effectiveness of using active learning to detect a dataset shift and we pay particular attention to the problem of the appearance of new classes that may not have been observed in the initial training set. To illustrate the proposed strategy, the Breaking Ties (BT) active sampling proposed by Luo et al. (2005) is used with a Linear Discriminant Analysis (LDA), which is a classifier widely used in real applications and also strongly prone to fail in case of covariate shift. Exploration of the data distribution through clustering is also used to cope with common situations, where one or several classes would not have been observed in the training set, but appear in the rest of the image. The proposed approach is tested on two urban and two agricultural remote sensing images, where the relevance of completing an existing training set with smartly selected pixels can be appreciated.

The remainder of the paper is organized as follows. Section 2 presents the problem of covariate shift and the proposed correction based on active learning. Section 3 details the data and the setup of the experiments discussed in section 4. Section 5 concludes the paper.

## 2. Covariate shift and active learning

This section briefly exposes the problem of covariate shift and converts it to a sampling problem. Active learning is then proposed as an alternative to fill the covariate shift gap. Finally, the problem of exploration is considered and a cluster-based heuristic is proposed to comply with the emergence of new, unexpected, classes.

### 2.1. The problem of covariate shift

Covariate shift is a common problem for any statistical model aiming at classifying a series of pixel vectors $\mathbf{x}$ into a series of land use classes $y$. The common assumption that the data are independent and identically distributed (*i.i.d*) usually does not hold for real applications, since the data distribution $p_{tr}(\mathbf{x})$ used for training the model only partially represents the

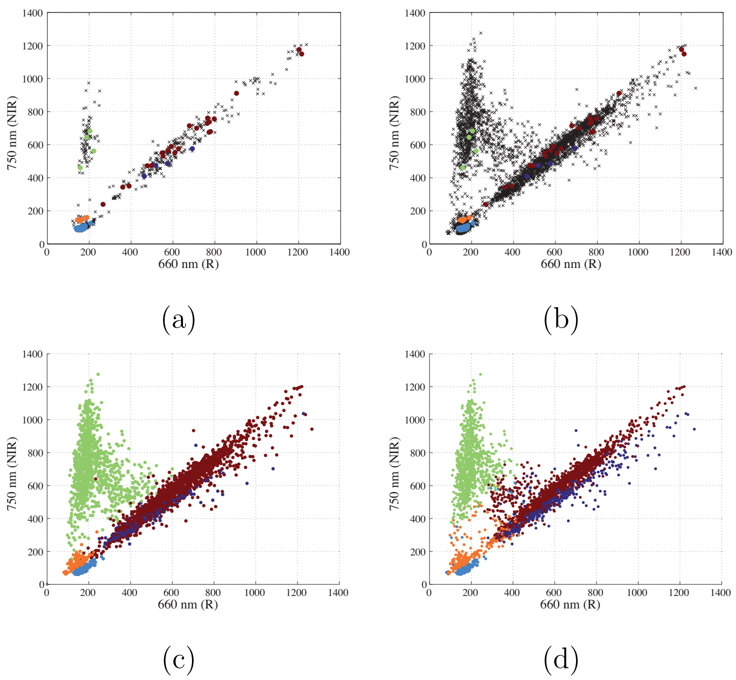



Figure 1: Dataset shift problem: (a) the training set (in color) is well suited to describe the unlabeled data (in black); (b) if using these training data to a larger amount of test data, the available training points become suboptimal with respect to the true labeling of the larger test set, shown in (c) subfigure corresponding to the labeling of bottom left part of Fig. 4; (d) a classifier such as LDA is thus prone to fail at classifying the test data. The data cloud is the one of the ROSIS image presented in Fig. 4.

true data distribution, that is represented by the test data distribution $p_{ts}(\mathbf{x})$. Nonetheless, it is a common assumption for machine learning algorithms to assume that test data follow the same joint probability distribution as the training data, i.e. $p(y|\mathbf{x})p_{tr}(\mathbf{x})$, where $y$ is the class label. Therefore, there is the risk that the new test data follow a slightly different distribution, which can be written for the same conditional distribution as $p(y|\mathbf{x})$, that $p_{tr}(\mathbf{x}) \approxeq p_{ts}(\mathbf{x})$. This situation is known as *covariate shift* and can result in a model that is optimal for a part of the data, but becomes sub-optimal if applied to the entire image. Figure 1 illustrates this phenomenon: a model trained on data coming from a part of a satellite image (the 'A' region of Fig. 4) can optimally describe the distribution of this sub-image, represented by the black crosses in Fig. 1a. When this same training set is used to describe the class distribution in the entire image (black crosses of Fig. 1b), the model fails because some areas of the feature space are not covered by this training set. Some of these areas were not present in the subset image, and represent the shift between the subset and the entire scene. Such a shift is related to differences in geometry that were not taken into account in the first place or to reflectances of the objects that were not covered by the available training set. When using LDA on this data, the true class memberships (shown in Fig. 1c) are not correctly represented in the outcome of the model (illustrated in Fig. 1d): the model built without adaptation models poorly at the interface between classes, thus resulting in an important decrease in the classification performance.

*2.2. Active learning to correct dataset shift*

Since the training and test distributions come from the same image, illumination conditions do not change and it is rather unlikely to find complex distortions between the two feature spaces: in this case, the shift is to be found in missing parts of the true data distribution (see Fig. 1a-b). Adapting the classifier trained on the subset to the entire image can be thus seen as efficiently finding the uncovered areas and sample useful pixels to classify them.

This is a typical setting for active learning algorithms (Cohn et al., 1994), which are algorithms aiming at finding efficient training sets to solve classification problems. For this particular problem, active learning results in a search for pixels enhancing the adaptation of the model to the rest of the image, i.e. refining the description of the boundaries between classes.

Active learning algorithms can be briefly summarized as follows (see

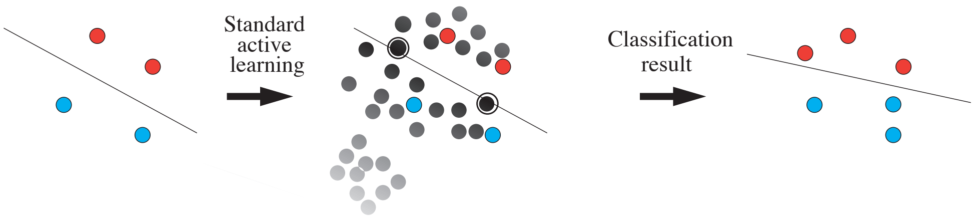



Figure 2: Uncertainty-based active learning algorithm general flowchart: (left) given an incomplete training set, (center) the unlabeled candidates are ranked according to a specific heuristic (represented by the grey tones attributed to the unlabeled pixels); (right) the candidates maximizing the heuristic are labeled and added to the training set.

Fig. 2): starting with a suboptimal training set composed by $n$ pixels $S = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$, an active learning algorithm exploits a ranking criterion, or *heuristic*, to rank all the $m$ unlabeled pixels $U = \{\mathbf{x}_j\}_{j=n+1}^{n+m}$ in order to select the most informative and add them to $S$. By so doing, the model is forced to focus on conflicting areas and to improve its generalization capabilities.

In this paper, the Breaking Ties heuristic proposed by Luo et al. (2005) is used: for each candidate, the two highest posterior class probabilities are subtracted, forming the ranking criterion that is exploited by the algorithm.

$$\hat{\mathbf{x}}^{\mathrm{BT}} = \arg\min_{\mathbf{x}_j \in U} \{\max_{\omega \in N} p(y_j^* = \omega|\mathbf{x}_j) - \max_{\omega \in N \setminus \omega^+} p(y_j^* = \omega|\mathbf{x}_j)\} \tag{1}$$

where $y_j^*$ is the class prediction for the pixel $\mathbf{x}_j$, $\omega \in N$ corresponds to one among the $N$ possible classes and $\omega^+ = \arg\max_{\omega \in N}\{p(y_j^* = \omega|\mathbf{x}_j)\}$ is the most probable class for pixel $\mathbf{x}_j$.

After ranking, the pixels maximizing Eq. (1) are then taken from the $U$ set, labeled by the user, and finally added to the current training set $S = \{S \cup \hat{\mathbf{x}}^{\mathrm{BT}}\}$. This heuristic uses the following intuition: the more a pixel shows a similar posterior probability between the two most probable classes, the more it is uncertain and thus capable of providing useful information if added to the training set. In previous experiments the BT approach has shown to be capable of providing good performance with remote sensing data (Copa et al., 2010).

### *2.3. On the need of an exploration-focused heuristic*

Using active queries to learn datasets seems an appealing solution for the classification of remote sensing data. However, the use of such models must be handled with care, since it relies on the quality of the initial training set (in our case, the available labeled pixels in the sub-image). If these pixels do not cover the entire distribution of the classes (which is reasonable in a covariate shift setting), there is also the possibility that a class will be ignored in the available training set. Consider again Fig. 2: in the central plot, there is a cluster of pixels in the bottom left part of the distribution. A traditional active learning algorithm, since it focuses on the uncertainty in the vicinity of the classification boundary only, will never check on the uncertainty of this region, since it is related to the data structure and not the current model uncertainty. As a consequence, this cluster will never be sampled by such an active algorithm. This may be problematic if this cluster corresponds to a new, unknown class. Approaches trying to constrain traditional heuristic to make them explore the feature space have been proposed in Ferecatu &

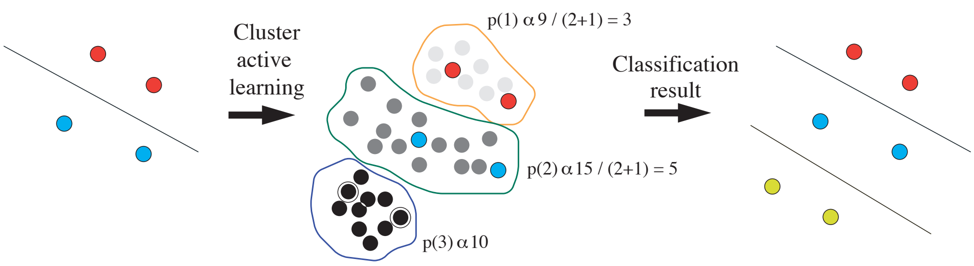



Figure 3: Cluster-based active learning algorithm general flowchart: (left) given an incomplete training set, (center) the unlabeled candidates are ranked according to the heuristic of Eq. (2) (in the computation, only the numerator is reported); (right) the candidates maximizing the heuristic are labeled and added to the training set, allowing the discovery of a third class.

Boujemaa (2007) and Tuia et al. (2009b), but they focus on the classification boundary and thus will also fail in this context.

Another view can be gained by using general data clustering, as in Xu et al. (2003) or Nguyen & Smeulders (2004): to cover the entire data distribution, we proceed to a pre-clustering of the image in a given number of clusters to decide whether there are some unexplored areas of the image. Contrary to these results, this process is not intended to create the initial training set, since a fair amount of labeled data are already available. Therefore, this knowledge about the availability of labeled samples can be used to direct sampling. We use a cost function aware of the presently available training samples, in the sense of Dasgupta & Hsu (2008). After clustering of the image in $k$ clusters using, for instance, $k$-means, pixels are iteratively chosen from the cluster $c_i$ with a probability proportional to the following heuristic:

$$p(c_i) \propto \frac{\frac{n_i}{l_i+1}}{\sum_{j=1}^{k} \frac{n_j}{l_j+1}} \tag{2}$$

where $n_i$ is the size of the cluster and $l_i$ is the number of labeled pixels already present in the cluster. In this way we sample from large and unseen clusters, where new classes are supposed to lie. This cluster-based strategy is summarized in Figure 3. After an iteration of this procedure, traditional active learning can be used to refine the classification boundaries defined.

## 3. Data and experimental setup

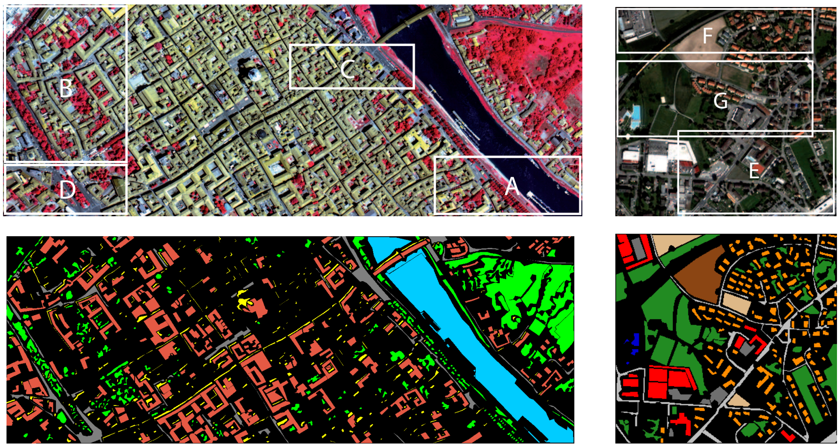
This section presents the dataset considered and details the setup of the experiments performed in Section 4.

### 3.1. Datasets

Two urban datasets at metric spatial resolution have been considered:

- The first data set is a 1.3 m resolution image of the city of Pavia (Italy), shown on the left side of Fig. 4. The image was taken by the airborne ROSIS-03 sensor (Licciardi et al., 2009). The image is $1400 \times 512$ pixels and has a spectral resolution from 0.43 to 0.86 $\mu$m divided into 102 spectral bands. The proposed approach has been tested on a 5-class problem, namely: Buildings, Roads, Water, Vegetation and Shadows. These classes of interest have been included in a labeled dataset of $206,009$ samples extracted by visual inspection.

- The second case study considers a 2.4 m resolution image of a suburb of the city of Zurich (Switzerland), shown on the right side of Fig. 4. The image has been acquired by the sensor on the QuickBird satellite and is a $329 \times 347$ pixel image with four spectral bands in the visible and near-infrared portions of the spectrum. A total of $43,398$ pixels have been labeled by visual inspection on the image with eight landuse classes have been selected for analysis (Residential, Commercial, Vegetation, Soil, Mixed soil / vegetation, Roads, Pools, Parkings). Note that several classes have very similar spectral signatures and, in order to differentiate them, contextual filters using mathematical morphology (Soille, 2004) with per-band opening and closing filters using spherical structure elements of 3 and 5 pixels diameter have been added to the dataset. This increases the dimensionality of the dataset from 4 to 20 features. These filters have been shown to have desirable proper-




ROSIS, Pavia QuickBird, Zurich

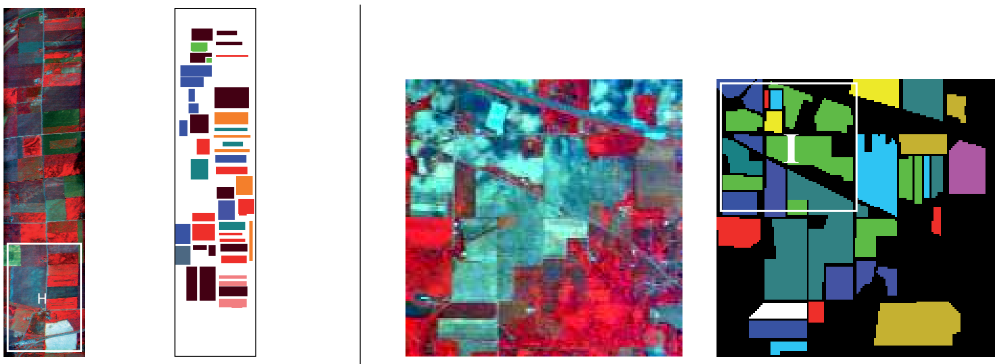
Figure 4: Top row: considered urban datasets. Areas marked by 'A' and 'B' (respectively 'E' and 'F') are the training areas of the experiments shown in Section 4. 'C' and 'D' (respectively 'G') areas are only used for graphics of an unseen area. Bottom row: available ground truth pixels.

ties when applied to urban VHR classification problems (Fauvel et al., 2008; Tuia et al., 2009a).

In addition, two agricultural datasets at medium spatial resolution have been considered[1]:

- The third dataset called Flightline C1 is a 12-bands multispectral image taken over Tippecanoe County, IN by the M7 scanner in June 1966 (Jackson & Landgrebe, 2001). The image is $949 \times 220$ pixels and contains 10 classes, mainly crop types. A ground survey of $70,847$ pixels

[1]Both datasets are available at https://engineering.purdue.edu/~biehl/MultiSpec/ hyperspectral.html




M7 scanner, Tippecanoe | AVIRIS, Indian Pines

Figure 5: Considered agricultural datasets and available ground truth pixels. Areas marked by ‘H’ and ‘I’ are the training areas of the experiments shown in Section 4.

has been used.

- The fourth dataset is the classical 220-bands AVIRIS image taken over Indiana’s Indian Pine test site in June 1992. The image is $145 \times 145$ pixels, contains 12 major crop types classes (with more than 100 labeled samples), and a total of $10,172$ labeled pixels. This image is a classical benchmark to validate model accuracy and constitutes a very challenging classification problem because of the strong mixture of the classes’ signatures and unbalanced number of labeled pixels per class. Before training the classifiers, we removed 20 noisy bands covering the region of water absorption and reduced the dimensionality to 6 features with PCA (accounting for 99.9% of data variance) to ensure correct estimation of the covariance matrix. As for the Zurich image, morphological opening and closing bands have been added to the extracted features. This is justified by the fact that the image has been taken shortly after plantation of the crops, thus showing class signatures that are, in fact, mixtures between soil and crops. Therefore, in order to achieve correct detection, contextual information must be added.

### *3.2. Experimental setup*

Experiments on urban areas use four training areas, each providing areas with increasing complexity in landcover.

A. this area covers all the classes present in the Pavia image. The shifts that need to be detected by the learning process are related to sampling in incomplete portions of the distribution. This first step can be considered as a classical active learning problem.

B. this area of the Pavia image lacks the class Water. In this example, we aim at discovering a major class (water covers a large part of the rest of the image) for a relatively easy classification problem. This experiment should reveal an inadequacy of traditional active learning since random sampling has a higher probability of finding this new class simply by chance.

E. this area of the Zurich image accounts for most of the classes except Water which for this image is a very marginal class. The aim of this experiment is to assess whether the cluster-based strategy is adapted to find small classes.

F. this experiment is the most complex for urban areas. The 'F' area of the Zurich image lacks two classes (Water and Bare Soil), one being major and the other marginal. In this case we want to assess the ability of the proposed approach to update the model to one with several new classes having different PDFs in the new image.

Regarding agricultural areas, we concentrate on the problem of discovering new classes. Two experiments with increasing complexity have been performed.

H. in this setting, the model is trained with a ground truth covering a small part of the image with reduced ground truth. Both major and marginal classes are missing. In particular, a major class is not reported in the initial ground survey ('Oats', in blue in Fig. 5), thus implying very poor performance of the model without samples from the new distribution.

I. this experiment is designed to test the algorithms proposed to discover classes with strongly overlapping spectra. As it has been mentioned above, the image was taken shortly after the crops were planted, so that each signature is not pure, rather a mixture between soil and crop, resulting in strongly overlapping classes. In this setting, three classes are unknown to the first model, 'Soybean-clean', 'Wheat' and 'Grass / pasture-mowed'. By the strong degree of mixture of the classes of this image with the unknown classes, this problem seems not to be suited for standard active learning algorithms.

For all experiments, 1) first the LDA classifier is optimized using 1000 pixels from the Pavia image (300 for the Zurich image, 300 for the Tippecanoe image, and 300 for the Indian Pines image) from the training sub-area and tested on the available ground truth in the same area. This experiment assesses the performance of the model for the subarea the training samples are drawn from. Afterwards, four experiments are added: 2) direct classification of the entire image with the same training data; 3) classification of the entire image using 1600 (1000, 600, and 2300) pixels randomly selected from the whole image; 4) starting with the 1000 (300, 300, and 300) pixels of the model locally optimal, sample 600 (700, 300, and 2000) pixels randomly; and 5) with the same initial set, actively sample 600 (700, 300, and 2000) pixels. Finally, 6) active sampling of 600 (700, 300, and 2000) pixels is applied after the clustering-based initial selection.

For active learning, BT active learning is implemented in MATLAB. Thirty (70, 30, and 100) iterations with 20 (10, 10, and 20) samples per iteration have been carried out. The differences in number of pixels per iteration and in the number of iterations are dictated by the different resolutions of the images and by the differences in complexity between the datasets respectively. Ten independent runs have been conducted to study stability of the solution with respect to initialization. Performance was evaluated in terms of overall accuracy (OA), Kappa statistic and standard deviations.

## 4. Results and discussion

This section presents and discusses the experimental results obtained by the proposed method on both the urban and the agricultural datasets.

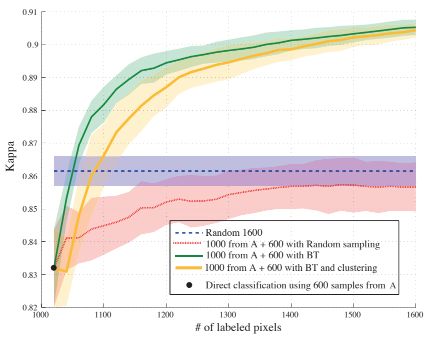
### *4.1. Urban data*

The first rows of Tab. 1 report the performance of the different strategies considered for the Pavia dataset by considering the patch 'A' as initial training area. When trained solely on the patch 'A', LDA performs perfectly when classifying that patch (OA = 98.42%), but fails on the entire image, where a decrease of about 12% in accuracy is observed (to 87.23%). A classifier trained on 1600 pixels randomly selected from the entire image can improve this result by approximatively 2% as does a random-based strategy sampling from the 1,000 initial samples. On the contrary, selecting the new pixels with active learning leads to an increase in performance of about 5% relative to the base classifier and 3% with respect to the experiment using 1,600 ran-

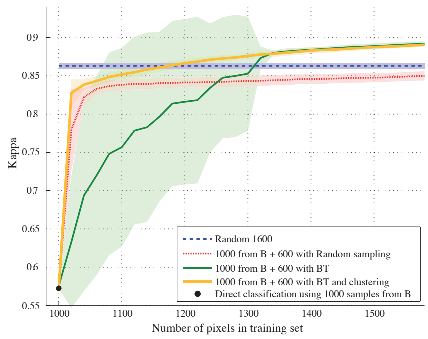
Table 1: Overall accuracy and Kappa statistic for the Pavia dataset. Iterative strategies are given at convergence. (* = Not comparable with the results of the other rows, different test sets).

| Training patch | Prediction area | # train (base) | # train (added) | Sampling strategy | OA $\mu$ | OA $\sigma$ | Kappa $\mu$ | Kappa $\sigma$ |
|---|---|---|---|---|---|---|---|---|
| A | A* | 1000 | – | – | 98.42 | 0.12 | 0.965 | 0.003 |
| | All image | 1000 | – | – | 87.23 | 0.70 | 0.827 | 0.009 |
| | All image | 1600 | – | – | 89.81 | 0.25 | 0.864 | 0.003 |
| | All image | 1000 | 600 | RS | 89.31 | 0.26 | 0.857 | 0.003 |
| | All image | 1000 | 600 | BT | **93.03** | **0.20** | **0.906** | **0.003** |
| | All image | 1000 | 600 | Cluster+BT | 92.97 | **0.17** | **0.905** | **0.002** |
| B | B* | 1000 | – | – | 85.81 | 0.74 | 0.767 | 0.012 |
| | All image | 1000 | – | – | 67.27 | 0.30 | 0.572 | 0.007 |
| | All image | 1600 | – | – | 89.78 | 0.28 | 0.863 | 0.004 |
| | All image | 1000 | 600 | RS | 88.83 | 0.46 | 0.850 | 0.006 |
| | All image | 1000 | 600 | BT | **91.98** | **0.25** | **0.892** | **0.003** |
| | All image | 1000 | 600 | Cluster+BT | 91.89 | 0.20 | 0.891 | 0.003 |

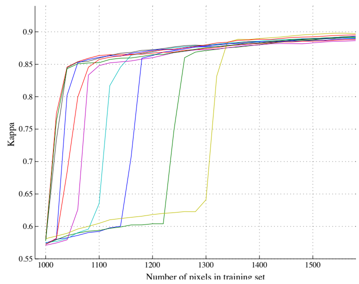
dom pixels. This approach reaches the best accuracy observed at 93.03% and 0.906 in terms of Kappa statistic. This is because the sampling is focused on the boundaries between classes where the shifts among distributions are more likely to occur. The curves of Fig. 6a show performance of the proposed methods as a function of the number of training samples. We note that the active learning process is faster to converge than it is the random selection process. In particular, 40 additive samples are sufficient for the standard BT method to reach the value of accuracy obtained by adding 600 random sam-




a)




b)




c)

Figure 6: Learning curves for the Pavia dataset. a) when using image patch ‘A’ for training set; b) when using image patch ‘B’ for training. c) Single runs composing the BT active learning curve (green curve in Fig. 6b).

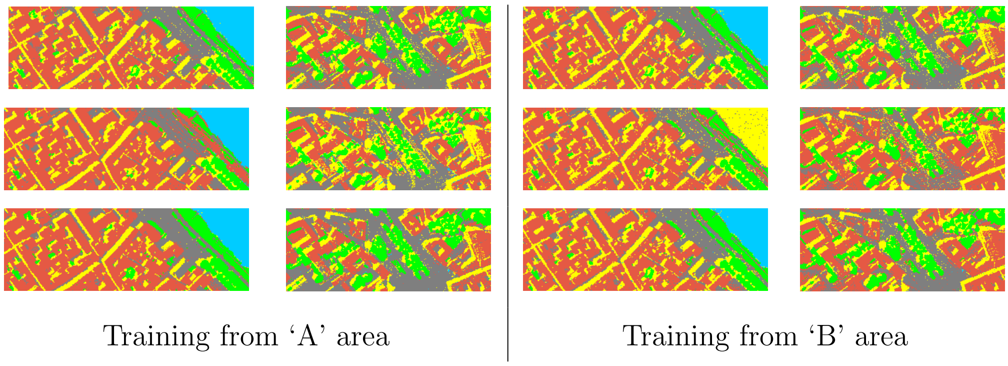



Figure 7: Classification maps for the Pavia dataset of regions ‘C’ and ‘D’ obtained when training LDA using pixels from regions (left) ‘A’ and (right) ‘B’. Top row illustrates the upper bound, where 1600 pixels randomly selected from the entire image. The middle row shows the experiment using the 1000 pixels only. The bottom row illustrates the results obtained by adding to these 1000 pixels 600 actively selected pixels from the rest of the image.

ples to the initial training set. Comparing orange and green curves, which are related to active sampling with and without clustering-based initialization respectively, we observe that the clustering strategy is not useful for this particular scenario. In fact, all the classes are already included in the initial training set, and so the initialization step tends to select samples that are not really important for better discriminating the different classes. In any case, a good improvement with respect to the random selection is preserved.

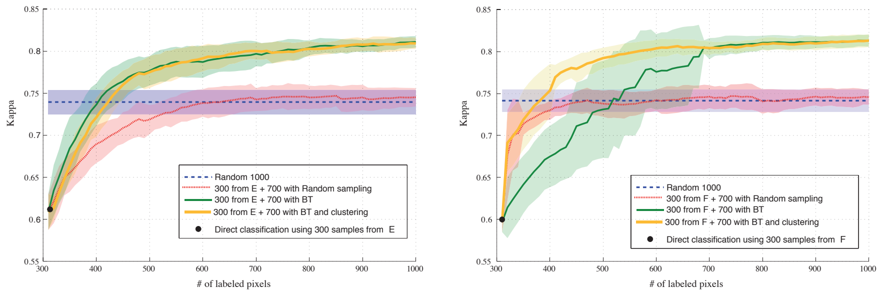
The results of the second experiment, in which the patch ‘B’ has been used to select the initial training set, are presented in the second part of Tab. 1. Because water pixels are not present in this patch, results show a strong decrease of LDA performance when applied to the entire image (from 85.81% to 67.27%). Sampling randomly from the entire image solves this problem, since the water class is well represented in the rest of the image and it is relatively easy to find by arbitrary sampling. Again, the active learning algorithm outperforms the others by 2-3% by focusing on the uncertain areas, resulting in an accuracy of 91.98% and 0.892 in Kappa. Regarding the curves in Fig. 6b, the active learning strategy is slower than the others to converge. The green curve in Fig. 6b is even worse than random selection in the first iterations. This can be explained by the plots of Fig. 1. If the water class is not found no area of uncertainty will be present for the class water and as a consequence such a class will never be sampled (unless by chance). The single runs generating the green curve in Fig. 6b are shown in Fig. 6c. The steep increase in accuracy for each run corresponds to the iteration where the water class is discovered. Applying the active learning after the clustering-based initialization, we have a fast convergence to optimal results avoiding overfitting, as illustrated by the orange curve in Fig. 6b. In this case, 180 additive samples are necessary to exceed the value of accuracy associated with the random selection.

These observations are confirmed by the maps shown in Fig. 7, in which a decrease of noisy classification patterns is obtained using the active learning strategy. Active strategies avoid sampling in already solved areas and thus reduce noisy classification results induced by sampling outliers.

Results obtained for the Zurich dataset confirm the considerations given for the Pavia image. For both patch 'E' and 'F' as initial training areas, active learning outperforms by about 5% the random selection method as described in Tab. 2. Once again the plots in Fig. 8 highlight the necessity of performing the initial selection with the clustering-based strategy when classes are missing in the initial training set. In particular, while this aspect

Table 2: Overall accuracy and Kappa statistic for the Zurich dataset. Iterative strategies are given at convergence. (* = Not comparable with the results of the other rows, different test sets).

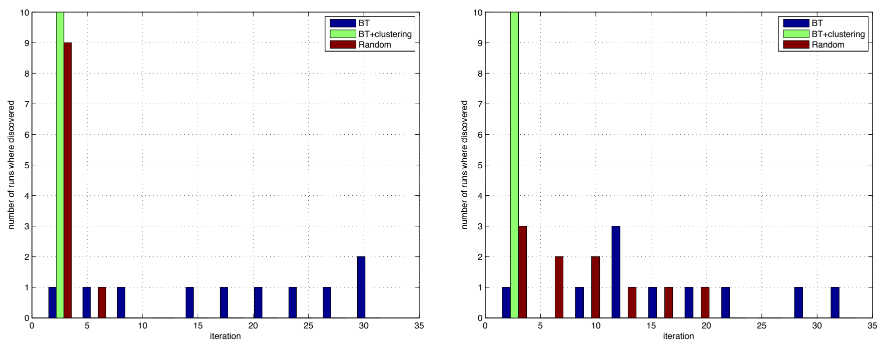
| Training patch | Prediction area | # train (base) | # train (added) | Sampling strategy | OA $\mu$ | OA $\sigma$ | Kappa $\mu$ | Kappa $\sigma$ |
|---|---|---|---|---|---|---|---|---|
| E | E* | 300 | – | – | 92.25 | 0.521 | 0.902 | 0.006 |
| | All image | 300 | – | – | 68.62 | 2.60 | 0.614 | 0.029 |
| | All image | 1000 | – | – | 79.48 | 1.23 | 0.743 | 0.014 |
| | All image | 300 | 700 | RS | 80.19 | 1.19 | 0.751 | 0.014 |
| | All image | 300 | 700 | BT | 85.07 | **0.58** | 0.809 | **0.007** |
| | All image | 300 | 700 | Cluster+BT | **85.35** | 0.68 | **0.813** | 0.008 |
| F | F* | 300 | – | – | 83.62 | 1.24 | 0.785 | 0.016 |
| | All image | 300 | – | – | 67.54 | 1.03 | 0.596 | 0.012 |
| | All image | 1000 | – | – | 78.87 | 1.49 | 0.736 | 0.017 |
| | All image | 300 | 700 | RS | 80.08 | 1.24 | 0.750 | 0.014 |
| | All image | 300 | 700 | BT | **85.25** | **0.67** | **0.812** | **0.008** |
| | All image | 300 | 700 | Cluster+BT | **85.24** | **0.68** | **0.812** | **0.008** |




a) b)

Figure 8: Learning curves for the Zurich dataset. a) when using image patch 'E' for training set; b) when using image patch 'F' for training.

is not crucial for the patch 'E', in which a single marginal class is not present initially, it becomes fundamental for the patch 'F', which lacks two classes, one major and the other marginal. Starting from the patch 'E', both strategies need 100 additional samples to reach the random sampling accuracy. For




a) b)

Figure 9: Number of runs for the Zurich dataset where classes missing in the image patch 'F' are discovered at each iteration. a) for major class Bare Soil; b) for marginal class Water.

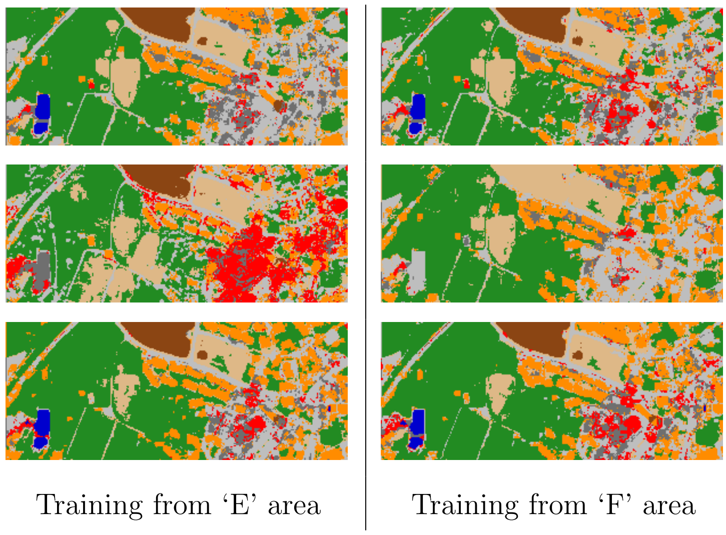



Figure 10: Classification maps for the Zurich dataset of the region ‘G’ obtained when training LDA using pixels from regions (left) ‘E’ and (right) ‘F’. Top row illustrates the upper bound, classifying 1000 pixels randomly selected from the entire image. The middle row shows the experiment using the 300 pixels only. The bottom row illustrates the results obtained by adding to these 300 pixels 700 actively selected pixels from the rest of the image.

patch ‘F’ only 80 instead of 220 samples are needed with clustering initialization relative to the traditional BT method. In the graph of Fig. 9, we report the number of runs where classes missing in patch ‘F’ are discovered at each iteration by the different methods proposed. For the class Bare Soil, shown in Fig. 9a, the initialization process is able to find it at the first iteration for all the ten runs considered. An identical behavior is obtained for the class Water (Fig. 9b), although the number of pixels of this class is very limited. An high probability of detection is verified for the random selection in the Bare Soil case, given the fact that it is easy to find this class by chance, while

Table 3: Overall accuracy and Kappa statistic for the (top) Tippecanoe and (bottom) Indian Pines datasets. Iterative strategies are given at convergence. (* = Not comparable with the results of the other rows, different test sets).

| Training patch | Prediction area | # train (base) | # train (added) | Sampling strategy | OA $\mu$ | OA $\sigma$ | Kappa $\mu$ | Kappa $\sigma$ |
|---|---|---|---|---|---|---|---|---|
| H | H* | 300 | – | – | 99.26 | 0.20 | 0.988 | 0.003 |
| | All image | 300 | – | – | 82.78 | 1.48 | 0.800 | 0.021 |
| | All image | 600 | – | – | 96.06 | 0.74 | 0.951 | 0.009 |
| | All image | 300 | 300 | RS | 96.04 | 0.53 | 0.951 | 0.006 |
| | All image | 300 | 300 | BT | 97.62 | 0.72 | 0.970 | 0.009 |
| | All image | 300 | 300 | Cluster+BT | **97.79** | **0.33** | **0.972** | **0.004** |
| I | I* | 300 | – | – | 72.52 | 2.21 | 0.671 | 0.026 |
| | All image | 300 | – | – | 43.70 | 0.80 | 0.365 | 0.009 |
| | All image | 2300 | – | – | 71.25 | 0.66 | 0.673 | 0.007 |
| | All image | 300 | 2000 | RS | 71.78 | **0.44** | 0.679 | **0.005** |
| | All image | 300 | 2000 | BT | 74.37 | 0.71 | 0.709 | 0.008 |
| | All image | 300 | 2000 | Cluster+BT | **74.69** | 1.07 | **0.713** | 0.012 |

poor performance is obtained for class Water. Finally, the traditional active sampling fails for both cases, where 30 iterations are needed to discover pixels of these classes in some runs. The final maps obtained for the Zurich image for the different proposed solutions are shown in Fig. 10.

### *4.2. Agricultural data*

Results obtained for the agricultural datasets are illustrated in Tab. 3 and corresponding Figs. 11 to 13.

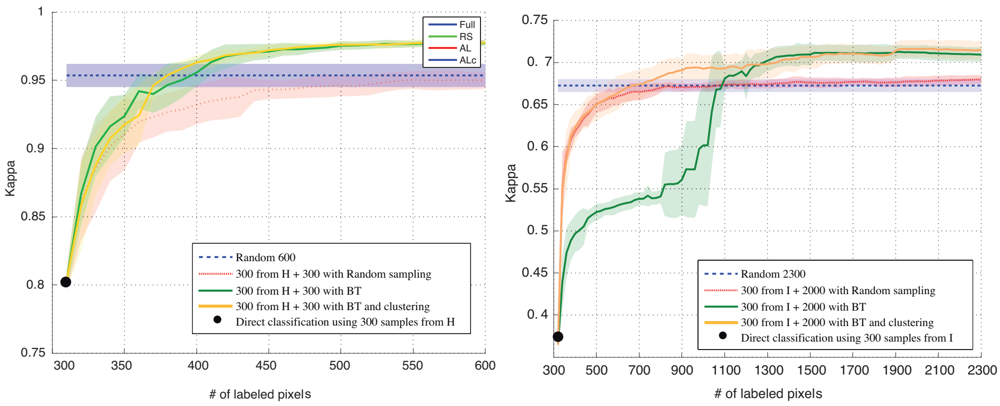



Figure 11: Learning curves for the agricultural datasets. left) Tippecanoe; right) Indian Pines.

At convergence, the results for the Tippecanoe image (training patch 'H') show an improvement with respect to random sampling by approximatively 2% and 0.02 in terms of accuracy and Kappa respectively, that is less spectacular than in the previous experiments. However, the learning rates show a strong divergence between the random and the active curves starting from iteration 3, when 360 samples are used for training (left side of Fig. 11). The similar behavior in the first two iterations is observed because the initial training set obviates most of the classes and then all the strategies perform well. Once the classification problem has become clearer, the active learning strategies can make difference, as shown in the figure. This behavior was already encountered and documented in Tuia et al. (2009b). As for the classification maps of Fig. 12, the active learning strategy returns a more desirable description of the class 'Rye' (in red), whose confusion with the class 'Soil' (in pink) is strongly diminished.

The last experiment considers the Indian Pines image. For this com-

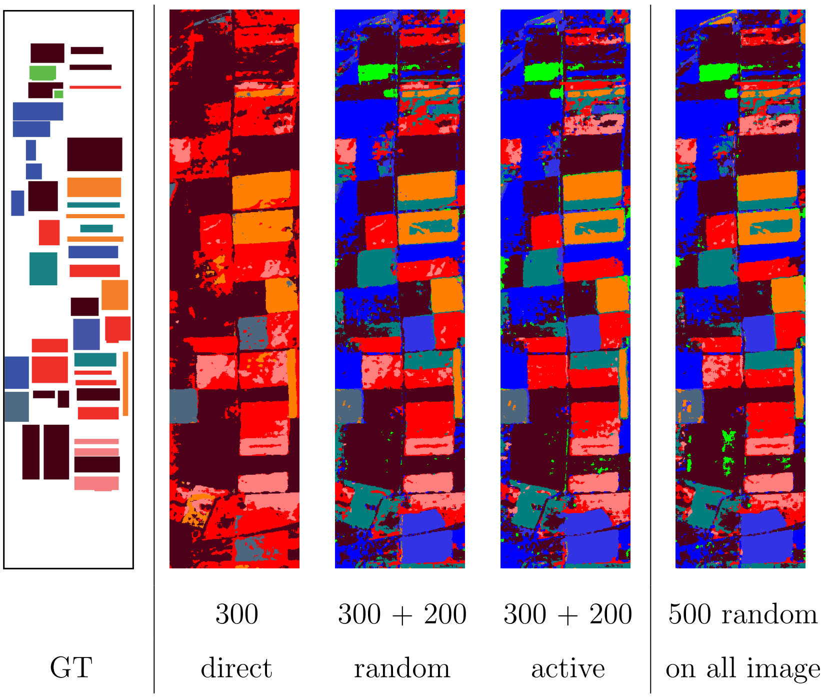



Figure 12: Classification maps for the Tippecanoe dataset using training information coming from the 'H' area after 10 iterations.

plex dataset, consisting classes showing strongly mixed signatures, the same behavior as in the urban dataset is observed (right side of Fig. 11): the traditional active learning strategy does not converge efficiently in the first iterations and is outperformed by random sampling. This again is due to the incapability of this strategy to discover new classes in highly overlapping problems. On the contrary, the proposed strategy considering pre-clustering performs efficiently, learns the global structure as efficiently as random sampling and outperforms it after 200 queries, reaching at convergence results higher by 3% in accuracy and 0.04 in Kappa. The classification maps obtained by this strategy, illustrated in Fig. 13, show a more homogeneous

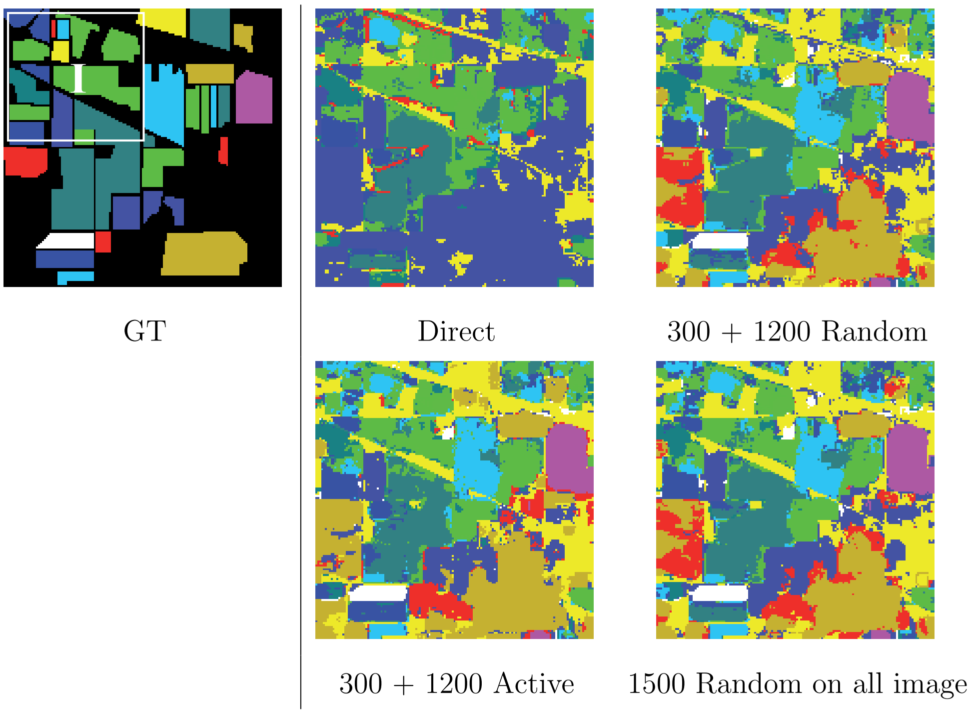



Figure 13: Classification maps for the Indian Pines dataset using training information coming from the 'I' area after 60 iterations.

result that the one obtained by random sampling.

## 5. Conclusion

In this paper, we have proposed a simple, but effective way to use active learning to solve the problem of dataset shift, which may occur when a classifier trained on a portion of the image is applied to the rest of the image. The experimental results obtained on hyperspectral and VHR datasets demonstrate good capability of the proposed method for selecting pixels that allow rapid convergence to an optimal solution. Moreover, the use of a clustering-based selection strategy allows us to discover new classes in case they have been omitted in the initial training set. Such strategies for optimal sampling guarantee signature extension and can be extended to a large variety of applications dealing with spectral data, as it is not dependent on the image characteristics of the data. Future research will explore these kinds of applications. An example could be the classification of Electrocardiographic signals, that has recently been tackled in Pasolli & Melgani (2010) using active learning techniques, but without considering issues related to covariate shift.

## Acknowledgments

This work has been partly supported by the Swiss National Science Foundation (grant no. PBLAP2-127713/1). The authors would like to acknowledge Prof. Paolo Gamba from the University of Pavia for providing the ROSIS data, Prof. Mikhail Kanevski from the University of Lausanne for providing the QuickBird data and Prof. M M. Crawford for providing the AVIRIS and M7 data.

## References

## References

Bickel, S., Brückner, M., & Scheffer, T. (2009). Discriminative learning under covariate shift. *J. Mach. Learn. Res.*, *10*, 2137–2155.

Bruzzone, L., & Fernandez-Prieto, D. (2001). Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, *39*, 456–460.

Bruzzone, L., & Marconcini, M. (2009). Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy. *IEEE Trans. Geosci. Remote Sens.*, *47*, 1108–1122.

Camps-Valls, G., & Bruzzone, L. (2009). *Kernel Methods for Remote Sensing Data Analysis*. NJ, USA: J. Wiley & Sons.

Cohn, D., Atlas, L., & R., L. (1994). Improving generalization with active learning. *Mach. Learn.*, *15*, 201–221.

Copa, L., Tuia, D., Volpi, M., & Kanevski, M. (2010). Unbiased query-by-bagging active learning for VHR image classification. In *Proceedings of the SPIE Remote Sensing Conference*. Toulouse, France.

Dasgupta, S., & Hsu, D. (2008). Hierarchical sampling for active learning. In *Intl. Conf. Mach. Learn. ICML* (pp. 208–215). Helsinki, Finland: ACM Press volume 307 of *ACM International Conference Proceeding Series*.

Fauvel, M., Benediktsson, J. A., Chanussot, J., & Sveinsson, J. R. (2008). Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. *IEEE Trans. Geosci. Remote Sens.*, *46*, 3804 – 3814.

Ferecatu, M., & Boujemaa, N. (2007). Interactive remote sensing image retrieval using active relevance feedback. *IEEE Trans. Geosci. Remote Sens.*, *45*, 818–826.

Fleming, M. D., Berkebile, J. S., & Hoffer, R. M. (1975). *Computer-aided analysis of LANDSAT-I MSS data: a comparison of three approaches,* *including a "Modified clustering" approach*. LARS information note 072475 Purdue University.

Foody, G. M., Boyd, D. S., & Cutler, M. E. J. (2003). Predictive relations of tropical forest biomass from landsat TM data and their transferability between regions. *Remote Sens. Environ.*, *85*, 463–474.

Gómez-Chova, L., Camps-Valls, G., Bruzzone, L., & Calpe-Maravilla, J. (2010). Mean map kernel methods for semisupervised cloud classification. *IEEE Trans. Geosci. Remote Sens.*, *48*, 207–220.

Gómez-Chova, L., Camps-Valls, G., Muñoz-Marí, J., & Calpe, J. (2008). Semi-supervised image classification with laplacian support vector machines. *IEEE Geosci. Remote Sens. Lett.*, *5*, 336–340.

Jackson, Q., & Landgrebe, D. (2001). An adaptive classifier design for high-dimensional data analysis with a limited training data set. *IEEE Trans. Geosci. Remote Sens.*, *39*, 2664–2679.

Jia, X., & Richards, J. A. (2002). Cluster-space representation for hyperspectral data classification. *IEEE Trans. Geosci. Remote Sens.*, *40 (3)*, 593–598.

Li, J., Bioucas-Dias, J., & Plaza, A. (2010a). Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Trans. Geosci. Remote Sens.*, *48*, 4085 –4098.

Li, Y., Kambara, H., Koike, Y., & Sugiyama, M. (2010b). Application of covariate shift adaptation techniques in brain computer interfaces. *IEEE Trans. Biomedic. Eng.*, *57*, 1318–1324.

Licciardi, G., Pacifici, F., Tuia, D., Prasad, S., West, T., Giacco, F., Inglada, J., Christophe, E., Chanussot, J., & Gamba, P. (2009). Decision fusion for the classification of hyperspectral data: Outcome of the 2008 GRS-S data fusion contest. *IEEE Trans. Geosci. Remote Sens.*, *47*, 3857–3865.

Liu, Q., Liao, X., & Carin, L. (2008). Detection of unexploded ordnance via efficient semisupervised and active learning. *IEEE Trans. Geosci. Remote Sens.*, *46*, 2558–2567.

Luo, T., Kramer, K., Goldgof, D. B., Hall, L. O., Samson, S., Remsen, A., & Hopkins, T. (2005). Active learning to recognize multiple types of plankton. *J. Mach. Learn. Res.*, *6*, 589–613.

Mitra, P., Uma Shankar, B., & Pal, S. (2004). Segmentation of multispectral remote sensing images using active support vector machines. *Pattern Recogn. Lett.*, *25*, 1067–1074.

Nguyen, H. T., & Smeulders, A. (2004). Active learning using pre-clustering. In *Intl. Conf. Mach. Learn. ICML* (pp. 623–630). Banff, Canada: ACM Press volume 69 of *ACM International Conference Proceeding Series*.

Olthof, I., Butson, C., & Fraser, R. (2005). Signature extension through space for northern landcover classification: A comparison of radiometric correction methods. *Remote Sens. Environ.*, *95*, 290–302.

Pasolli, E., & Melgani, F. (2010). Active learning methods for electrocardiographic signal classification. *IEEE Trans. Information Tech. in Biomedicine*, *14*, 1405–1416.

Pasolli, E., Melgani, F., & Bazi, Y. (in press). SVM active learning through significance space construction. *IEEE Geosci. Remote Sens. Lett.*, .

Patra, S., & Bruzzone, L. (in press). A fast cluster-assumption based active-learning technique for classification of remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, .

Pax-Lenney, M., Woodcock, C. E., Macomber, S. A., Gopal, S., & Song, C. (2001). Forest mapping with a generalized classifier and landsat TM data. *Remote Sens. Environ.*, *77*, 241–250.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). *Dataset Shift in Machine Learning*. MIT Press.

Rajan, S., Ghosh, J., & Crawford, M. (2006). Exploiting class hierarchy for knowledge transfer in hyperspectral data. *IEEE Trans. Geosci. Remote Sens.*, *44*, 3408–3417.

Rajan, S., Ghosh, J., & Crawford, M. (2008). An active learning approach to hyperspectral data classification. *IEEE Trans. Geosci. Remote Sens.*, *46*, 1231–1242.

Schweikert, G., Widmer, C., Schölkopf, B., & Rätsch, G. (2008). An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

Soille, P. (2004). *Morphological image analysis*. Berlin-Heidelberg: Springer-Verlag.

Sugiyama, M., Krauledat, M., & Müller, K. R. (2007). Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.*, *8*, 985–1005.

Tuia, D., & Camps-Valls, G. (2009). Semi-supervised remote sensing image classification with cluster kernels. *IEEE Geosci. Remote Sens. Lett.*, *6*, 224–228.

Tuia, D., Pacifici, F., Kanevski, M., & Emery, W. (2009a). Classification of very high spatial resolution imagery using mathematical morphology and support vector machines. *IEEE Trans. Geosci. Remote Sens.*, *47*, 3866–3879.

Tuia, D., Ratle, F., Pacifici, F., Kanevski, M., & Emery, W. (2009b). Active learning methods for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.*, *47*, 2218–2232.

Woodcock, C. E., Macomber, S. A., Pax-Lenney, M. P., & Cohen, W. B. (2001). Monitoring large areas for forest change using landsat: Generalization across space, time and landsat sensors. *Remote Sens. Environ.*, *78*, 194–203.

Xu, Z., Yu, K., Tresp, V., Xu, X., & Wang, J. (2003). Representative sampling for text classification using support vector machines. In *25th European Conf. on Information Retrieval Research* (pp. 393–407).

Yang, J., Yan, R., & Hauptmann, A. G. (2007). Cross-domain video concept detection using adaptive SVMs. In *15th international conference on Multimedia* (pp. 188 – 197). Ausburg, Germany: ACM Press.