# A survey of active learning algorithms for supervised remote sensing image classification

Devis Tuia, *Member, IEEE*, Michele Volpi, *Student Member, IEEE*, Loris Copa, Mikhail Kanevski, Jordi Muñoz-Marí



**Abstract**

**This is the pre-acceptance version, to read the final version published in 2011 in the IEEE Journal of Selected Topics in Signal Processing (IEEE JSTSP), please go to: 10.1109/JSTSP.2011.2139193**

Defining an efficient training set is one of the most delicate phases for the success of remote sensing image classification routines. The complexity of the problem, the limited temporal and financial resources, as well as the high intraclass variance can make an algorithm fail if it is trained with a suboptimal dataset. Active learning aims at building efficient training sets by iteratively improving the model performance through sampling. A user-defined heuristic ranks the unlabeled pixels according to a function of the uncertainty of their class membership and then the user is asked to provide labels for the most uncertain pixels. This paper reviews and tests the main families of active learning algorithms: committee, large margin and posterior probability-based. For each of them, the most recent advances in the remote sensing community are discussed and some heuristics are detailed and tested. Several challenging remote sensing scenarios are considered, including very high spatial resolution and hyperspectral image classification. Finally, guidelines for choosing the good architecture are provided for new and/or unexperienced user.

Manuscript received April 2010;

This work has been supported by the Swiss National Science Foundation (grants no. 200021-126505 and PBLAP2-127713/1), and by the Spanish Ministry of Education and Science under projects TEC2009- 13696, AYA2008-05965-C04-03, and CONSOLIDER/CSD2007-00018.

DT and JMM are with the Image Processing Laboratory, University of València, València, Spain. C/ Cat. A. Escardino. 46980 Paterna, València, Spain. Email: {devis.tuia,jordi}@uv.es, http://ipl.uv.es, Phone: +34 963544021, Fax: +34 963544353

MV and MK are with the Institute of Geomatics and Analysis of Risk, University of Lausanne, Lausanne, Switzerland. Email: {michele.volpi,mikhail.kanevski}@unil.ch, http://www.unil.ch/igar, Phone: +4121-6923546, Fax: +4121-6923535

LC was with the Institute of Geomatics and Analysis of Risk, University of Lausanne, Lausanne, Switzerland. He is now with SARMAP SA, Switzerland. Email: loris.copa@sarmap.ch, Phone +41 916009365

**Index Terms**

Image classification, Active learning, Training set definition, SVM, VHR, Hyperspectral.

## I. INTRODUCTION

Nowadays, the recourse to statistical learning models [1] is a common practice for remote sensing data users; models such as Support Vector Machines (SVM, [2], [3]) or neural networks [4] are considered as state of the art algorithms for the classification of landuse using new generation satellite imagery [5]. Applications of such models to very high spatial [6]–[8] or spectral [9]–[11] resolution have proven their efficiency for handling remote sensing data.

However, the performances of supervised algorithms strongly depend on the representativeness of the data used to train the classifier [12]. This constraint makes the generation of an appropriate training set a difficult and expensive task requiring extensive manual analysis of the image. This is usually done by visual inspection of the scene or by field surveys and successive labeling of each sample.

In the case of field surveys, which is usual for medium resolution, hyperspectral or SAR images, the discovery of a new label is expensive – both in terms of time and money – because it involves terrain campaigns. Therefore, there is a limit to the number of pixels that can be acquired. For this reason, compact and informative training sets are needed.

In the case of visual inspection or photo-interpretation, more common in VHR imagery, it is easier to collect data samples, since the labeling can be done directly on the image. However, the labeling is often done by mass selection on screen and several neighboring pixels carrying the same information are included. As a consequence, the training set is highly redundant. Such a redundancy considerably slows down the training phase of the model. Moreover, the inclusion of noisy pixels may result in a wrong representation of the class statistics, which may lead to poor classification performances and/or overfitting [13]. For these reasons, a training set built by photointerpretation should also be kept as small as possible and focused on those pixels effectively improving the performance of the model.

Summing up, besides being small, a desirable training set must be constructed in a smart way, meaning it must represent correctly the class boundaries by sampling discriminative pixels. This is particularly critical in very high spatial and spectral resolution image classification, which deal with large and/or complex features spaces using limited training information only [14].

In the machine learning literature this approach to sampling is known as *active learning*. The leading idea is that a classifier trained on a small set of well-chosen examples can perform as well as a classifier trained on a larger number of randomly chosen examples, while it is computationally cheaper [15]–[17]. Active learning focuses on the interaction between the user and the classifier. In other words, the model returns to the user the pixels whose classification outcome is the most uncertain. After accurate labeling by the user, pixels are included into the training set in order to reinforce the model. This way, the model is optimized on well-chosen difficult examples, maximizing its generalization capabilities.

The active learning framework has demonstrated its effectiveness when applied to large datasets needing accurate selection of examples [18]. This is suitable for remote sensing applications, where the number of pixels among which the search is performed is large and manual definition is - as stated above - redundant and time consuming. As a consequence, active learning algorithms gain an increasing interest in remote sensing image processing and several approaches have been proposed to solve image classification tasks. This paper presents the general framework of active learning and reviews some of the methods that have been proposed in remote sensing literature. Note that this survey only covers remote sensing application of active learning principles: for a general introduction and survey of the most recent developments in the machine learning community, please refer to [19], [20].

The remainder of the paper is organized as follows: Section II presents the general framework of active learning and the families of methods that will be detailed in Sections III to V, as well as the references to specific methods. Section VI presents the datasets considered in the experiments. Section VII compares the different approaches numerically. Section VIII gives an overview and guidelines for potential users. Section IX concludes the paper.

## II. Active learning: concepts and definitions

Let $X = \{\mathbf{x}_i, y_i\}_{i=1}^{l}$ be a training set of labeled samples, with $\mathbf{x}_i \in \mathcal{X}$ and $y_i = \{1, ..., N\}$. $\mathcal{X}$ is the $d$-dimensional input space $\in \mathbb{R}^d$. Let also $U = \{\mathbf{x}_i\}_{i=l+1}^{l+u} \in \mathcal{X}$, with $u \gg l$ be the set of unlabeled pixels to be sampled, or the *pool of candidates*.

Active learning algorithms are iterative sampling schemes, where a classification model is adapted regularly by feeding it with new labeled pixels corresponding to the ones that are most beneficial for the improvement of the model performance. These pixels are usually found in the

**Algorithm 1** General active learning algorithm

**Inputs**

- Initial training set $X^\epsilon = \{\mathbf{x}_i, y_i\}_{i=1}^{l}$ ($X \in \mathcal{X}$, $\epsilon = 1$).
- Pool of candidates $U^\epsilon = \{\mathbf{x}_i\}_{i=l+1}^{l+u}$ ($U \in \mathcal{X}$, $\epsilon = 1$).
- Number of pixels $q$ to add at each iteration (defining the batch of selected pixels $S$).

**repeat**
  Train a model with current training set $X^\epsilon$.
  **for** each candidate in $U^\epsilon$ **do**
    Evaluate a user-defined *heuristic*
  **end for**
  Rank the candidates in $U^\epsilon$ according to the score of the heuristic
  Select the $q$ most interesting pixels. $S^\epsilon = \{\mathbf{x}_k\}_{k=1}^{q}$
  The user assigns a label to the selected pixels. $S^\epsilon = \{\mathbf{x}_k, y_k\}_{k=1}^{q}$
  Add the batch to the training set $X^{\epsilon+1} = X^\epsilon \cup S^\epsilon$.
  Remove the batch from the pool of candidates $U^{\epsilon+1} = U^\epsilon \backslash S^\epsilon$
  $\epsilon = \epsilon + 1$
**until** a stopping criterion is met.

areas of *uncertainty* of the model and their inclusion in the training set forces the model to solve the regions of low confidence. For a given iteration $\epsilon$, the algorithm selects from the pool $U^\epsilon$ the $q$ candidates that will at the same time maximize the gain in performance and reduce the uncertainty of the model when added to the current training set $X^\epsilon$. Once the batch of pixels $S^\epsilon = \{\mathbf{x}_m\}_{m=1}^{q} \subset U$ has been selected, it is labeled by the user, i.e. the labels $\{y_m\}_{m=1}^{q}$ are discovered. Finally, the set $S^\epsilon$ is both added to the current training set ($X^{\epsilon+1} = X^\epsilon \cup S^\epsilon$) and removed from the pool ($U^{\epsilon+1} = U^\epsilon \backslash S^\epsilon$). The process is iterated until a stopping criterion is met. Algorithm 1 summarizes the active selection process. From now on, the iteration index $\epsilon$ will be omitted in order to ease notation.

An active learning process requires interaction between the user and the model: the first

provides the labeled information and the knowledge about the desired classes, while the latter provides both its own interpretation of the distribution of the classes and the most relevant pixels that are needed in order to solve the discrepancies encountered. This point is crucial for the success of an active learning algorithm: the machine needs a strategy to rank the pixels in the pool $U$. These strategies, or *heuristics*, differentiate the algorithms proposed in the next sections and can be divided into three main families [21]:

1 - *Committee*-based heuristics (Section III)
2 - *Large margin*-based heuristics (Section IV)
3 - *Posterior probability*-based heuristics (Section V)

A last family of active learning heuristics, the *cluster*-based, has recently being proposed in the community [22]: cluster-based heuristics aim at pruning a hierarchical clustering tree until the resulting clusters are consistent with the labels provided by the user. Therefore, these strategies rely on an unsupervised model, rather than on a predictive model. Since the aim of these heuristics is different form that of the other families presented, they will not be detailed in this survey.

## III. Committee based active learning

The first family of active learning methods quantifies the uncertainty of a pixel by considering a committee of learners [23], [24]. Each member of the committee exploits different hypotheses about the classification problem and consequently labels the pixels in the pool of candidates. The algorithm then selects the samples showing maximal disagreement between the different classification models in the committee. To limit computational complexity, heuristics based on multiple classifier systems have been proposed in machine learning literature. In [25], methods based on boosting and bagging are proposed in this sense for binary classification only. In [26], results obtained by query-by-boosting and query-by-bagging are compared using several batch datasets showing excellent performance of the heuristics proposed. Methods of this family have the advantage to be applicable to any kind of model or combination of models. In the remote sensing community, committee-based approaches to active learning have been proposed exploiting two types of uncertainty: first, committees varying the pixels members have been considered in the query-by-bagging heuristic [21], [27]. Then, committees based on subsets of

the feature space available have been presented in Di and Crawford [28]. The next two sections present the algorithms proposed in these papers.

### *A. Normalized entropy query-by-bagging (nEQB)*

In the implementations of [25], bagging [29] is proposed to build the committee: first, $k$ training sets built on a draw with replacement of the original data are defined. These draws account for a part of the available labeled pixels only. Then, each set is used to train a classifier and to predict the $u$ labels of the candidates. At the end of the procedure, $k$ labelings are provided for each candidate pixel $\mathbf{x}_i \in U$. In [21], the entropy $H^{\text{BAG}}$ of the distribution of the predictions is used as heuristic. In [27], this measure has been subsequently normalized in order to bound it with respect to the number of classes predicted by the committee and avoid hot spots of the value of uncertainty in regions where several classes overlap. The *normalized entropy query-by-bagging* heuristic can be stated as follows:

$$\hat{\mathbf{x}}^{n\text{EQB}} = \arg\max_{\mathbf{x}_i \in U} \left\{ \frac{H^{\text{BAG}}(\mathbf{x}_i)}{\log(N_i)} \right\} \tag{1}$$

where

$$H^{\text{BAG}}(\mathbf{x}_i) = -\sum_{\omega=1}^{N_i} p^{\text{BAG}}(y_i^* = \omega|\mathbf{x}_i) \log\left[p^{\text{BAG}}(y_i^* = \omega|\mathbf{x}_i)\right] \tag{2}$$

$$\text{where} \qquad p^{\text{BAG}}(y_i^* = \omega|\mathbf{x}_i) = \frac{\sum_{m=1}^{k} \delta(y_{i,m}^*, \omega)}{\sum_{m=1}^{k} \sum_{j=1}^{N_i} \delta(y_{i,m}^*, \omega_j)}$$

$H^{\text{BAG}}(\mathbf{x}_i)$ is an empirical measure of entropy, $y_i^*$ is the prediction for the pixel $\mathbf{x}_i$ and $p^{\text{BAG}}(y_i^* = \omega|\mathbf{x}_i)$ is the observed probability to have the class $\omega$ predicted using the training set $X$ by the committee of $k$ models for the sample $\mathbf{x}_i$. $N_i$ is the number of classes predicted for pixel $\mathbf{x}_i$ , with $1 \leq N_i \leq N$. The $\delta(y_{i,m}^*, \omega)$ operator returns the value 1 if the classifier using the $m$-th bag classifies the sample $\mathbf{x}_i$ into class $\omega$ and 0 otherwise. Entropy maximization gives a naturally multiclass heuristic. A candidate for which all the classifiers in the committee agree is associated with null entropy and its inclusion in the training set does not bring additional information. On the contrary, a candidate with maximum disagreement between the classifiers results in maximum entropy, and its inclusion will be highly beneficial.

### *B. Adaptive maximum disagreement (AMD)*

When confronted to high dimensional data, it may be relevant to construct the committee by splitting the feature space into a number of subsets, or *views* [30]. Di and Crawford [28] exploit this principle to generate different views of a hyperspectral image on the basis of the block-diagonal structure of the covariance matrix. By generating views corresponding to the different blocks, independent classifications of the same pixel can be generated and an entropy-based heuristic can be used similarly to $n$EQB.

Given a partition of the $d$-dimensional input space into $V$ disjoint views accounting for data subsets $\mathbf{x}^v$ such that $\bigcup_{v=1}^{V} \mathbf{x}^v = \mathbf{x}$, the 'Adaptive maximum disagreement ' (AMD) heuristic selects candidates according to the following rule:

$$\hat{\mathbf{x}}^{\text{AMD}} = \arg\max_{\mathbf{x}_i \in U} H^{\text{MV}}(\mathbf{x}_i) \tag{3}$$

where the multiview entropy $H^{\text{MV}}$ is assessed over the predictions of classifiers using a specific view $v$:

$$H^{\text{MV}}(\mathbf{x}_i) = -\sum_{\omega=1}^{N_i} p^{\text{MV}}(y^*_{i,v} = \omega|\mathbf{x}^v_i) \log\left[p^{\text{MV}}(y^*_{i,v} = \omega|\mathbf{x}^v_i)\right] \tag{4}$$

$$\text{where} \qquad p^{\text{MV}}(y^*_i = \omega|\mathbf{x}^v_i) = \frac{\sum_{v=1}^{V} W^{\epsilon-1}(v,\omega)\delta(y^*_{i,v},\omega)}{\sum_{v=1}^{V}\sum_{j=1}^{N_i} W^{\epsilon-1}(v,\omega)}$$

where the $\delta(y^*_{i,v},\omega)$ operator returns the value 1 if the classifier using the view $v$ classifies the sample $y_i$ into class $\omega$ and 0 otherwise. $\mathbf{W}^{\epsilon-1}$ is a $N \times V$ weighting matrix incorporating the abilities of discrimination between the views in the different classes. At each iteration, $\mathbf{W}^{\epsilon-1}$ is updated on the basis of the true labels of the pixels sampled at iteration $\epsilon - 1$:

$$W^{\epsilon}(v,\omega) = W^{\epsilon-1}(v,\omega) + \delta(y_{i,v},\omega), \quad \forall i \in S \tag{5}$$

and its columns are normalized to a unitary sum. This matrix weights the confidence of each view to predict a given class. In [28], the selection is done on a subset of $U$ containing the candidate pixels maximizing the uncertainty, which are the pixels for which the committee has predicted the highest number of classes. This way, the computational load of the algorithm is reduced.

## IV. LARGE MARGIN BASED ACTIVE LEARNING

The second family of methods is specific to margin-based classifiers. Methods such as SVM are naturally good base methods for active learning: the distance to the separating hyperplane, that is the absolute value of the decision function without the sign operator, is a straightforward way of estimating the classifier confidence on an unseen sample. Let's first consider a binary problem: the distance of a sample $\mathbf{x}_i$ from the SVM hyperplane is given by

$$f(\mathbf{x}_i) = \sum_{j=1}^{n} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i) + b \tag{6}$$

where $K(\mathbf{x}_j, \mathbf{x}_i)$ is a kernel, which defines the similarity between the candidate $\mathbf{x}_i$ and the support vectors $\mathbf{x}_j$, which are the pixels showing non zero $\alpha_j$ coefficients. The labels $y_j$ of the support vectors are $+1$ for samples of the positive class and $-1$ for those on the negative. For additional information, see SVM literature in [2], [3].

This evaluation of the distance is the base ingredient of almost all large margin heuristics. Roughly speaking, these heuristics use the intuition that a sample away from the decision boundary (with a high $f(\mathbf{x}_i)$) has a high confidence about its class assignment and is thus not interesting for future sampling.

Since SVM rely on a sparse representation of the data, large margin based heuristics aim at finding the pixels in $U$ that are most likely to receive a non-zero $\alpha_i$ weight if added to $X$. In other words, the points more likely to become support vectors are the ones lying within the margin of the current model [31]. The heuristic taking advantage of this property is called margin sampling (MS) [32], [33]. Recent modifications of MS, aiming at minimizing the risk of selecting points that will not become support vectors, can be found in [34], [35]. MS is the most studied active learning algorithm in remote sensing. Its first application can be found in [36]. Modifications of the MS heuristic have been proposed in [37]. Later on, since no cross-information among the samples is considered in the MS, the questions of diversity in batches of samples have been considered in [21], [38], [39]. The next sections present the MS heuristic and subsequent modifications proposed in order to enhance diversity when selecting batches of samples.

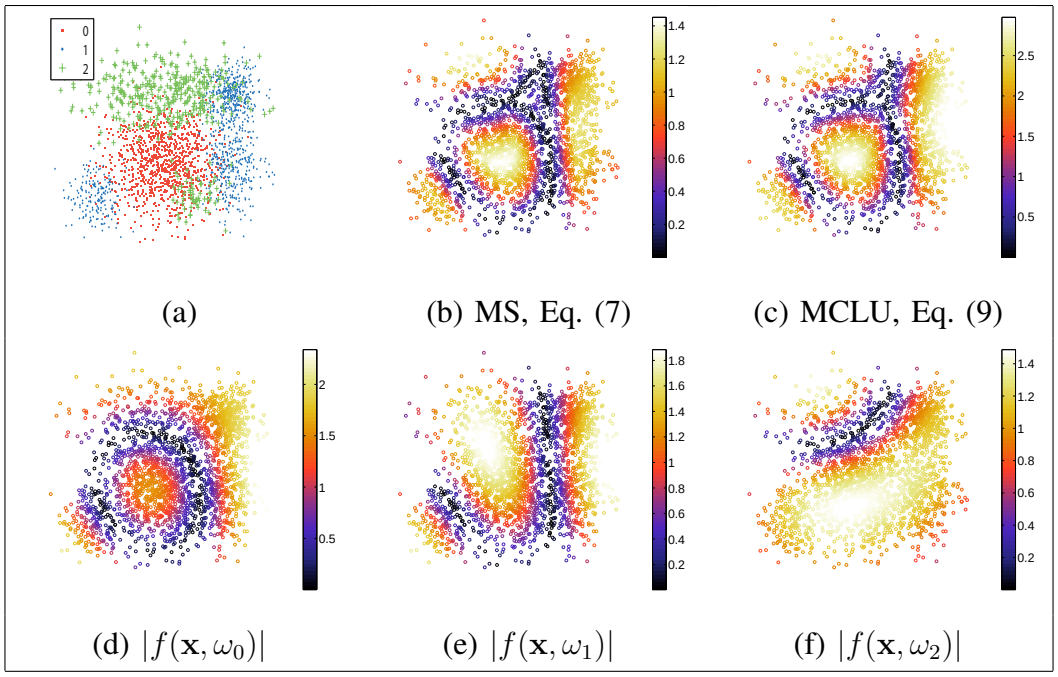



Fig. 1. Large margin heuristics for a three classes toy example represented in subfigure (a). The color intensity represents the distance from the hyperplane, ranging from black (on the boundary) to white (maximal distance): (b) MS heuristic; (c) MCLU heuristic; areas in black are the areas of maximal uncertainty, minimizing Eq. (7) or Eq. (9) respectively. Bottom row: absolute values of per-class distances (d)-(f).

### *A. Margin sampling (MS)*

As stated above, margin sampling takes advantage of SVM geometrical properties, and in particular of the fact that unbounded support vectors are labeled examples that lie on the margin with a decision function value of exactly one [2], [3]. Consider the pool of candidates of Fig. 1(a) referring to a three classes toy problem. In a multiclass one-against-all setting, the distance to each hyperplane is represented by Figs. 1(d-f). The 'margin sampling' (MS) heuristic performs sampling of the candidates by minimizing Eq. (7):

$$\hat{\mathbf{x}}^{\text{MS}} = \arg\min_{\mathbf{x}_i \in U} \left\{ \min_{\omega} |f(\mathbf{x}_i, \omega)| \right\} \tag{7}$$

where $f(\mathbf{x}_i, \omega)$ is the distance of the sample to the hyperplane defined for class $\omega$ in a one-against-all setting for multiclass problems. The MS heuristic for the toy problem is reported in Fig. 1(b). MS heuristic can be found in the literature under the names of 'most ambiguous' [38], 'binary level uncertainty' [39] or $\text{SVM}_{\text{SIMPLE}}$ [28].

### *B. Multiclass level uncertainty (MCLU)*

In [39], the idea of margin sampling is extended to multiclass uncertainty (see [40]). Instead of dealing with the most uncertain class of the SVM, the 'multiclass level uncertainty' (MCLU) considers the difference between the distance to the margin for the two most probable classes.

$$\hat{\mathbf{x}}^{\text{MCLU}} = \arg\min_{\mathbf{x}_i \in U} \left\{ f(\mathbf{x}_i)^{\text{MC}} \right\} \tag{8}$$

$$\text{where} \qquad f(\mathbf{x}_i)^{\text{MC}} = \max_{\omega \in N} |f(\mathbf{x}_i, \omega)| - \max_{\omega \in N \setminus \omega^+} |f(\mathbf{x}_i, \omega)| \tag{9}$$

where $\omega^+$ is the class showing maximal confidence, i.e. the argument of the first term of Eq. (9) showing maximal $f(\mathbf{x}_i)^{\text{MC}}$. A high value of this criterion corresponds to samples assigned with high certainty to the most confident class, while a small value represents unreliable classification. Fig. 1(c) illustrates the heuristic in comparison to MS. Although they are very similar, MCLU performs better in the area where the three classes mix, in the top-right area of the feature space: in this area, MCLU returns maximal uncertainty as it is evaluated on the basis of all the per-class decision values, while MS returns an uncertainty slightly lower than on the two-classes boundaries.

### *C. Significance space construction (SSC)*

In [41], instead of using the distance to the hyperplane as a measure of uncertainty, the support vector coefficients are used to convert the multiclass classification problem into a binary support vector detection problem. In the 'significance space construction' (SSC) heuristic, the training samples related to support vector coefficients are used to define a second classification function $f(\mathbf{x})^{\text{SSC}}$, where training pixels with $\alpha_j > 0$ (the support vectors) are classified against training pixels with $\alpha_j = 0$. Once applied to the pool of candidates, this second classifier predicts which pixels are likely to become support vectors.

$$\hat{\mathbf{x}}^{\text{SSC}} = \arg_{\mathbf{x}_i \in U} f(\mathbf{x}_i)^{\text{SSC}} > 0 \tag{10}$$

Once the candidates more likely to become support vectors have been highlighted, a random selection among them is done.

### *D. On the need for a diversity criterion*

In applicative scenarios, diversity among samples [42] is highly desirable. Diversity concerns the capability of the model to reject candidates that rank well according to the heuristic, but are redundant among each other. Diversity has been studied extensively for margin-based heuristics, where the base margin sampling heuristic is constrained using a measure of diversity between the candidates (see Algorithm 2).

The first heuristic proposing explicit diversity in remote sensing is found in [38], where the margin sampling heuristic is constrained with a measure of the angle between candidates in feature space. This heuristic, called 'most ambiguous and orthogonal' (MAO) is iterative: starting from the samples selected by MS, $U^{\text{MS}} \subset U$, this heuristic iteratively chooses the samples minimizing the highest values between the candidates list and the samples already included in

---

**Algorithm 2** General diversity based heuristic (for a single iteration)

---

**Inputs**

- Current training set $X^\epsilon = \{\mathbf{x}_i, y_i\}_{i=1}^{l}$ ($X \in \mathcal{X}$).
- Subset of the pool of candidates minimizing Eq. (7) or (8) $U^* = \{\mathbf{x}_i\}$ ($U^* \in \mathcal{X}$ and $U^* \subset U^\epsilon$).
- Number of pixels $q$ to add at each iteration (defining the batch of selected pixels $S$).

1: Add the pixel minimizing Eq. (7) or (9) to $S$.
2: **repeat**
3: Compute the user defined diversity criterion between pixels in $U^*$ and in $S$ (with MAO, cSV or ABD).
4: Select the pixel $\mathbf{x}_D$ maximizing diversity with current batch.
5: Add $\mathbf{x}_D$ to current batch $S = S \cup \mathbf{x}_D$.
6: Remove $\mathbf{x}_D$ to current list of cadidates $U^* = U^* \setminus \mathbf{x}_D$.
7: **until** batch $S$ contains $q$ elements.
8: The user labels the selected pixels. $S = \{\mathbf{x}_k, y_k\}_{k=1}^{q}$
9: Add the batch to the training set $X^{\epsilon+1} = X^\epsilon \cup S$.
10: Remove the batch from the complete pool of candidates $U^{\epsilon+1} = U^\epsilon \backslash S$

---

the batch $S$. For a single iteration, this can be resumed as:

$$\hat{\mathbf{x}}^{\text{MAO}} = \arg\min_{\mathbf{x}_i \in U^{\text{MS}}} \left\{ \max_{\mathbf{x}_j \in S} K(\mathbf{x}_i, \mathbf{x}_j) \right\} \tag{11}$$

In [39], the MAO criterion is combined with the MCLU uncertainty estimation in the 'multi-class level uncertainty - angle-based diversity' (MCLU-ABD) heuristic. Te selection is performed among a subset of $U$ maximizing the MCLU criterion. Moreover, the author generalize the MAO heuristic to any type of kernels by including normalization in feature space.

$$\hat{\mathbf{x}}^{\text{MCLU-ABD}} = \arg\min_{\mathbf{x}_i \in U^{\text{MCLU}}} \left\{ \lambda f(\mathbf{x}_i)^{\text{MC}} + \right. \tag{12}$$

$$\left. (1-\lambda) \max_{\mathbf{x}_j \in S} \frac{K(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{K(\mathbf{x}_i, \mathbf{x}_i) K(\mathbf{x}_j, \mathbf{x}_j)}} \right\} \tag{13}$$

where $f(\mathbf{x}_i)^{\text{MC}}$ is the multiclass uncertainty function defined by Eq. (9).

In [21], the diversity of the chosen candidates is enforced by constraining the MS solution to pixels associated to different closest support vectors. This approach ensures a certain degree of diversification in the MS heuristic, by dividing the margin in the feature space as a function of the geometrical distribution of the support vectors. Compared to the previously presented heuristics, this approach has the advantage of ensuring diversity with respect to the current model, but does not guarantee diversity of the samples between each other (since two close samples can be associated to different support vectors).

$$\hat{\mathbf{x}}^{\text{cSV}} = \arg\min_{\mathbf{x}_i \in U^{\text{MS}}} \left\{ |f(\mathbf{x}_i, \omega)| \;\middle|\; cSV_i \notin cSV_\theta \right\} \tag{14}$$

where $\theta = [1, \ldots, q-1]$ are the indices of the already selected candidates and $cSV$ is the set of selected closest support vectors.

Finally, diversity can be ensured using clustering in the feature space. In [39], kernel $k$-means [43]–[45] was used to cluster the samples selected by MCLU and select diverse batches. After partitioning the $U^{\text{MCLU}}$ set into $q$ clusters with kernel $k$-means, the 'multiclass level uncertainty - enhanced cluster based diversity (MCLU-ECBD)' selects a single pixel per cluster, minimizing the following query function:

$$\hat{\mathbf{x}}^{\text{MCLU-ECBD}} = \arg\min_{\mathbf{x}_i \in c_m} \left\{ f(\mathbf{x}_i)^{\text{MC}} \right\}, \quad m = [1, ...q], \mathbf{x}_i \in U^{\text{MCLU}} \tag{15}$$

where $c_m$ is one among the $q$ clusters defined with kernel $k$-means.

In [46], a hierarchical extension of this principle is proposed to exclude from the selected batch the pixels more likely to become bounded support vectors. This way, the redundancy affecting samples close to each other in the feature space among different iterations is controlled along with the maximization of the informativeness of each pixel. In the 'informative hierarchical margin cluster sampling' (hMCS-i), a dataset composed by i) a subset of the pool of candidates optimizing the MCLU criterion ($U^{\text{MCLU}}$) and ii) the bounded support vectors sampled at the previous iteration, is iteratively partitioned in a binary way. The partitioning always considers the biggest current cluster found and continues until $q$ clusters not containing a bounded support vector are found. Once the $q$ clusters have been defined, a search among the candidates falling in these $q$ clusters is performed.

$$\hat{\mathbf{x}}^{\text{hMCS-i}} = \arg\min_{\mathbf{x}_i \in c_m} \left\{ f(\mathbf{x}_i)^{\text{MC}} \right\}, \quad m = [1, ..., q | n_{c_m}^{bSV} = 0], \mathbf{x}_i \in U^{\text{MCLU}} \tag{16}$$

## V. Posterior probability based active learning

The third class of methods uses the estimation of posterior probabilities of class membership (i.e. $p(y|\mathbf{x})$) to rank the candidates. The posterior probability gives an idea of the confidence of the class assignment (which is usually done by maximizing it over all the possible classes): by considering the change on the overall posterior distribution or the per-class distribution for each candidate, these heuristics use these probability estimates to focus sampling in uncertain areas. This section details two heuristics, the KL-max and the Breaking ties.

### *A. KL-max*

The first idea is to sample the pixels whose inclusion in the training set would maximize the changes in the posterior distribution. An application of these methods can be found in [47], where the heuristic maximizes the Kullback-Leibler divergence between the distributions before and after adding the candidate. In remote sensing, a probabilistic method based on this strategy and using a Maximum Likelihood classifier can be found in [48]. In this setting, each candidate is removed from $U$ and it is included in the training set with the label maximizing

its posterior probability. The Kullbach-Leibler divergence KL is then computed between the posterior distributions of the models with and without the candidate. After computing this measure for all candidates, the pixel maximizing the following is chosen:

$$\hat{\mathbf{x}}^{\text{KL-max}} = \tag{17}$$
$$= \arg\max_{\mathbf{x}_i \in U} \left\{ \sum_{\omega \in N} \frac{1}{(u-1)} \text{KL}\Big(p^+(\omega|\mathbf{x}) \Big|\Big| p(\omega|\mathbf{x})\Big) p(y_i^* = \omega|\mathbf{x}_i) \right\}$$

where

$$\text{KL}\Big(p^+(\omega|\mathbf{x}) \Big|\Big| p(\omega|\mathbf{x})\Big) = \sum_{\mathbf{x}_j \in U \setminus \mathbf{x}_i} p^+(y_j^* = \omega|\mathbf{x}_j) \log \frac{p^+(y_j^* = \omega|\mathbf{x}_j)}{p(y_j^* = \omega|\mathbf{x}_j)} \tag{18}$$

and $p^+(\omega|\mathbf{x})$ is the posterior distribution for class $\omega$ and pixel $\mathbf{x}$, estimated using the increased training set $X^+ = [X, (\mathbf{x}_i, y_i^*)]$, where $y_i^*$ is the class maximizing the posterior probability. Recently, the authors of [49] extended this approach, proposing to use boosting to weight pixels that were previously selected, but were no longer relevant to the current classifier. These heuristics are useful when used with classifiers with small computational cost: since each iteration implies to train $u+1$ models, this type of heuristics is hardly applicable with computationally demanding methods as SVM. Moreover, a selection of batches of pixels is not possible.

### *B. Breaking ties (BT)*

Another strategy, closer to the idea of EQB presented in Section III-A, consists of building a heuristic exploiting the conditional probability of predicting a given label $p(y_i^* = \omega|\mathbf{x}_i)$ for each candidate $\mathbf{x}_i \in U$. In this case, note that the predictions for the single candidates $y_i^* = \arg\max_{\omega \in N} f(\mathbf{x}_i, \omega)$ are used. Such estimates are provided by several methods, including probabilistic neural networks or maximum likelihood classifiers. A possibility to obtain posterior probabilities from SVM outputs[1] is to use Platt's estimation [50]. In this case, the per-class posterior probability is assessed fitting a sigmoid function to the SVM decision function [50]:

$$p(y_i^* = \omega|\mathbf{x}_i) = \frac{1}{1 + e^{(Af(\mathbf{x}_i, \omega) + B)}} \tag{19}$$

where $A$ and $B$ are parameters to be estimated (for details, see [50]). Once the posterior probabilities are obtained, it is possible to assess the uncertainty of the class membership for

[1]Even though they are not posterior probabilities from a Bayesian point of view

each candidate in a direct way. In this case the heuristic choses candidates showing a near uniform probability of belonging to each class, i.e. $p(y_i^* = \omega|\mathbf{x}_i) = 1/N,\ \forall \omega \in N$.

The 'Breaking ties' (BT) heuristic for a binary problem relies on the smallest difference of the posterior probabilities for each sample [51]. In a multi-class setting, this reciprocity can still be confirmed and used, since independently from the number of classes $N$, the difference between the two highest probabilities can be indicative of the way an example is handled by the classifier. When the two highest probabilities are close ("on a tie"), the classifier's confidence is the lowest. The BT heuristic can thus be formulated as:

$$\hat{\mathbf{x}}^{\text{BT}} = \arg\min_{\mathbf{x}_i \in U} \left\{ \max_{\omega \in N} \left\{ p(y_i^* = \omega|\mathbf{x}) \right\} - \max_{\omega \in N \setminus \omega^+} \left\{ p(y_i^* = \omega|\mathbf{x}) \right\} \right\} \tag{20}$$

where $\omega^+$ is the class showing maximal probability, i.e. the argument of the first term of Eq. (20). By comparing Eq. (8) with Eq. (20), it is clear that the link between BT and the MCLU heuristic when using SVM classifiers (see Section IV-B) is really strong.

## VI. Datasets and setup

This Section details the datasets considered and the setup of the experiments performed.

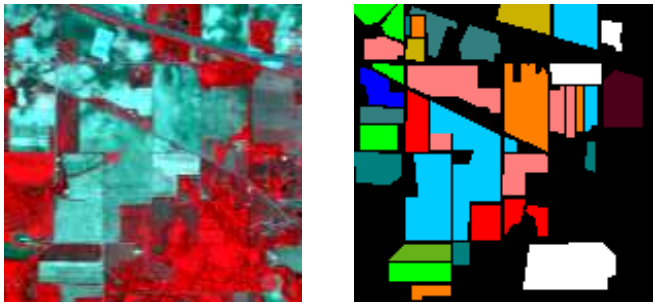
### A. Datasets

Active heuristics have been tested on three challenging remote sensing classification scenarios (Fig. 2), whose data distributions are detailed in Fig. 3.

*1) Hyperspectral VHR:* the two top rows of Fig. 2 show a hyperspectral 1.3m spatial resolution image of the city of Pavia (Italy) taken by the airborne ROSIS-03 optical sensor [52]. The image consists of 102 spectral bands of size $(1400 \times 512)$ pixels with a spectral coverage ranging from 0.43 to 0.86 $\mu$m. $5$ classes of interest (Buildings, Roads, Water, Vegetation and Shadows) have been selected and a labeled dataset of 206'009 pixels has been extracted by visual inspection. Among the available pixels, 20'000 have been used for the training set $X$ and candidate set $U$. Ten independent experiments have been performed, starting with $5 \times 5 = 25$ labeled pixels ($5$ per class) in $X$ and the remaining pixels in $U$. When using LDA, 150 pixels ($30$ per class) have been included in the starting set. The higher number of starting training pixels used for LDA is justified by the requirements of the model ($n$ must be greater than the dimensionality

of the data). In each experiment, 80‘000 randomly selected pixels have been used to test the generalization capabilities of the heuristics.

ROSIS Pavia



AVIRIS Indian Pines

QuickBird Zurich

Fig. 2. Images considered in the experiments: (top) ROSIS image of the city of Pavia, Italy (bands $[56-31-6]$ and corresponding ground survey); (middle) AVIRIS Indian Pines hyperspectral data (bands $[40-30-20]$ and corresponding ground survey); (bottom) QuickBird multispectral image of a suburb of the city of Zurich, Switzerland (bands $[3-2-1]$ and corresponding ground survey).

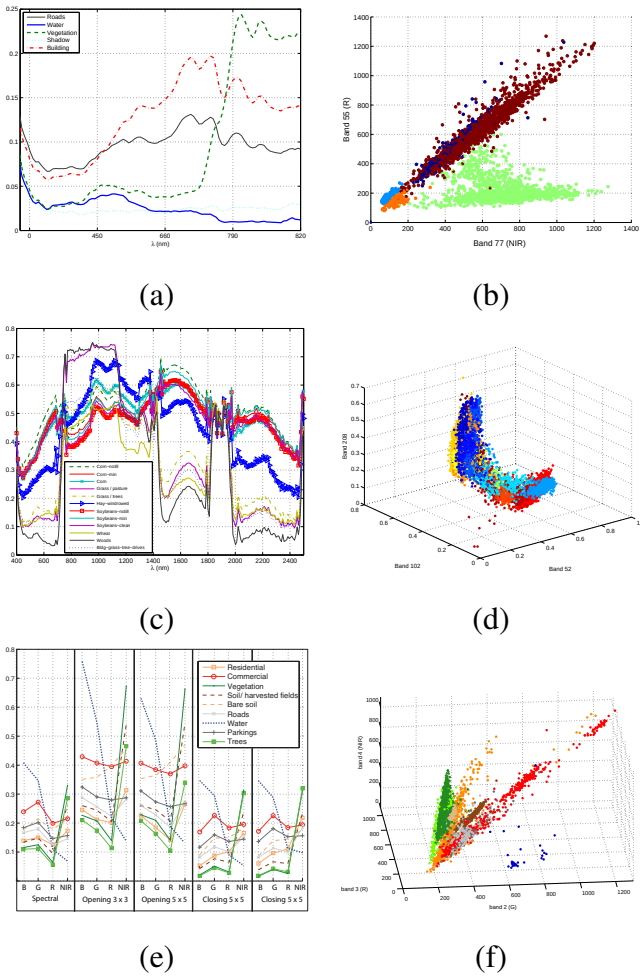



Fig. 3. Data distribution of the three images considered. First row: ROSIS image of Pavia: (a) mean spectral profiles; (b) example of data manifold in bands 55 (Red) and 77 (Near infrared). Middle row: AVIRIS Indian Pines: (c) mean spectral profiles; (d) example of data manifold in bands 52, 102 and 208. Bottom row: Zurich QuickBird: (e) mean spectral profiles; (f) data manifold in bands 2 (G), 3 (R) and 4 (NIR).

The data distribution of the five classes is illustrated in the first row of Fig. 3: from the mean spectra (Fig. 3(a)) the classes are well distinguished and separable with the sole spectral information and the resulting data manifold (Fig. 3(b)) shows a data distribution which can be handled by most linear and non linear models.

*2) Hyperspectral MR:* the second dataset, illustrated in the second row of Fig. 2, is a 220-bands AVIRIS image taken over Indiana's Indian Pine test site in June 1992 [53]. The image is $145 \times 145$ pixels, contains 16 classes representing different crops, and a total of 10'366 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 class signatures. Twenty water absorption channels were removed prior to analysis . In the experiments, classes with less than 100 labeled pixels were removed, resulting thus in a 12 classes classification problem with 10‘171 labeled pixels (see the ground truth pixels in Fig. 2). Among the available labeled pixels, 7‘000 were used for the $X$ and $U$ sets. Each experiment starts with $5 \times 12 = 60$ pixels (5 per class). As for the previous image, the remaining 3‘171 pixels have been used to test the generalization capabilities.

Visualization of the spectral mean profiles and of the data manifold (second row of Fig. 3) illustrates a completely different situation with respect to the previous image: high nonlinearity and strongly overlapping classes characterize this dataset. Therefore, linear classifiers do not perform well on this dataset and will not be considered in the experiments.

*3) Multispectral VHR:* The third image, illustrated in the last row of Fig. 2, is a 4-bands QuickBird scene of a suburb of the city of Zurich (Switzerland) taken in 2002. The image is $329 \times 347$ pixels with a spatial resolution of 2.4m. Nine classes of interest have been extracted by careful visual inspection, namely Residential buildings, Commercial buildings, Trees, Vegetation, Harvested fields, Bare soil, Roads, Parking lots and Water. Since some of the classes to be separated are of landuse and have very similar responses in the spectral domain (see, for instance, the residential and commercial buildings, or roads and parking lots), 16 contextual bands, extracted using opening (8) and closing (8) morphological operators (see [8]), have been added to the four spectral bands, resulting in a 20-dimensional dataset. As for the Pavia dataset, 20‘000 pixels have been extracted for the $X$ and $U$ sets. Each experiment starts with $5 \times 9 = 45$ pixels ($5$ per class). The complexity of this third dataset is confirmed by both the spectra and the manifold illustrated in the bottom row of Fig. 3. Strong overlaps between the asphalt and the soil classes are observed, which is also confirmed by the similarity between the spectral profiles. However, the spatial features added improve the differentiation of the classes (see, for instance, the opening features for the vegetation classes and the closing features for the asphalt classes).

### *B. Experimental setup*

In the experiments, SVM classifiers with RBF kernel and LDA classifiers have been considered for the experiments. When using SVM, free parameters have been optimized by 5-fold cross validation optimizing an accuracy criterion. The active learning algorithms have been run in two

settings, adding $N+5$ and $N+20$ pixels per iteration. To reach convergence, $70$ ($40$ in the case $N+20$) iterations have been executed for the first image, $100$ ($50$) for the second and $80$ ($50$) for the third. $n$EQB has been run with committees of $7$ models using $75\%$ of the available training data. For the experiments using LDA, $40$ (20) iterations have been performed and $n$EQB using $12$ models and $85\%$ of the data have been used.

An upper bound on the classification accuracy has been computed by considering a model trained on the whole $X \cup U$ set ('Standard SVM/LDA' hereafter). The lower bound on performance has been considered by assessing a model using an increasing training set, randomly adding the same number of pixels at each epoch ('Random Sampling' hereafter).

Each heuristic has been run ten times with different initial training sets. All the graphics report mean and standard deviation of the ten experiments.

## VII. Numerical results

In this section, some of the heuristics presented are compared on the three datasets presented above. The experiments do not aim at defining which heuristic is best, since they respond unequally well to different data architectures. Rather, it attempts to illustrate the strengths and weaknesses of the different families of methods and to help the user in selecting the methodology that will perform best depending on the characteristics of the problem at hand. The heuristics studied are the following: $n$EQB, MS, MCLU, MCLU-ABD and BT. Their comparison with Random sampling (RS), the base learner (Standard SVM/LDA) and between each other will show the main differences among the active learning architectures presented.

Figure 4 compares a heuristic for each family presented, when using SVM classifiers. In general, MS performs better than the two other families. This is expected, since MS ranks the candidates directly using the SVM decision function without further estimations: the slightly inferior performances of the $n$EQB and the BT algorithms are mainly due to the small size of the initial training set, which does not allow these criteria to perform optimally. $n$EQB uses too small bags of training pixels and BT cannot estimate the posterior probabilities correctly, because the fit of Platt's probabilities is dependent on the number of samples used. As a consequence, the performances of these two heuristics in the first iterations is similar to random sampling, a behavior already observed in [21]. Summarizing, when using SVMs, the most logical choice among the families seems to be a large margin heuristic.

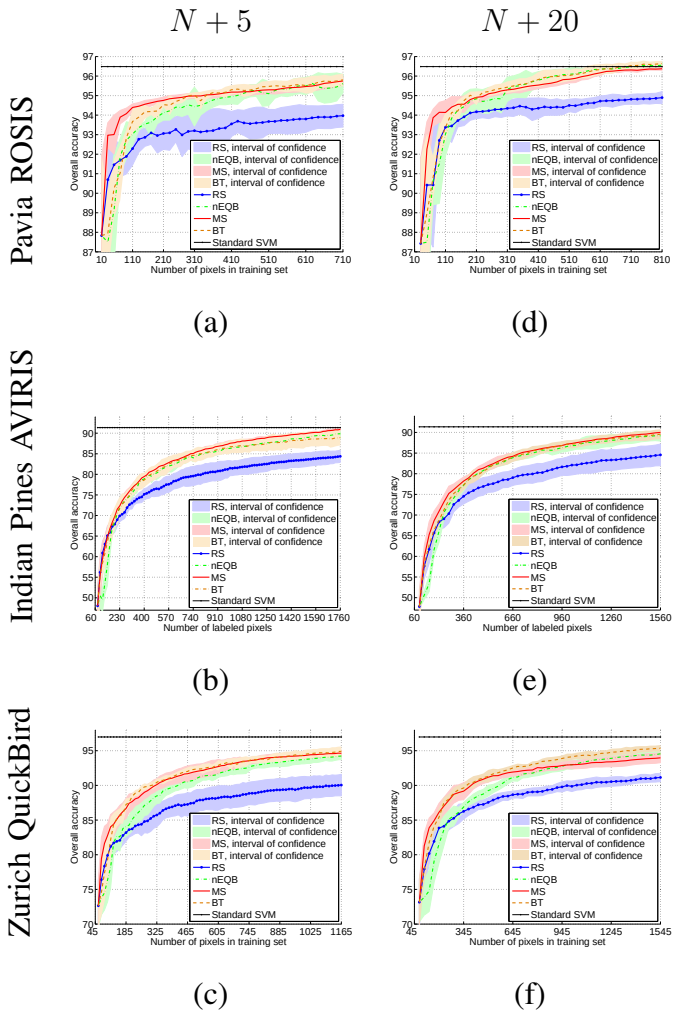



Fig. 4. The three families of heuristics trained with SVMs (RS = Random Sampling).

Regarding this family, Figs. 5 and 6 illustrate two concepts regarding the two stages of large margin heuristics: the uncertainty and diversity criteria. Figures 5 compares the MS and MCLU criteria and shows that both describe the uncertainty of the candidates in a similar way. Therefore, both can be used for efficient active learning. The use of a diversity criterion seems to slightly improve the quality of the results (Fig. 6): except for the AVIRIS image – well known for the high degree of mixture of its spectral signatures – a spectral diversity criterion such as MCLU-ABD efficiently increases performances with little added computational cost. None of the solutions obtained with the inclusion of the diversity criterion degrade the ones relying on the uncertainty assumption only.

As stated above, other heuristics must be used for other classifiers. Figure 7 compares the $n$EQB and BT heuristics applied to the Pavia image using LDA.

In this case, both heuristics perform similarly and show a very interesting behavior: active sampling helps the LDA to estimate the subspace that better separates data. In fact, when sampling randomly, noise and outliers can make the estimation of the Fisher's ratio biased, resulting in a suboptimal linear combination of variables in the decision function. Sampling and assigning correct labels to the pixels returned by these heuristics help estimating the correct *per-class* extent (covariance) and position (mean). From 330 pixels up, the standard LDA result is improved by the active learning training sets, providing more harmonious solutions that allow

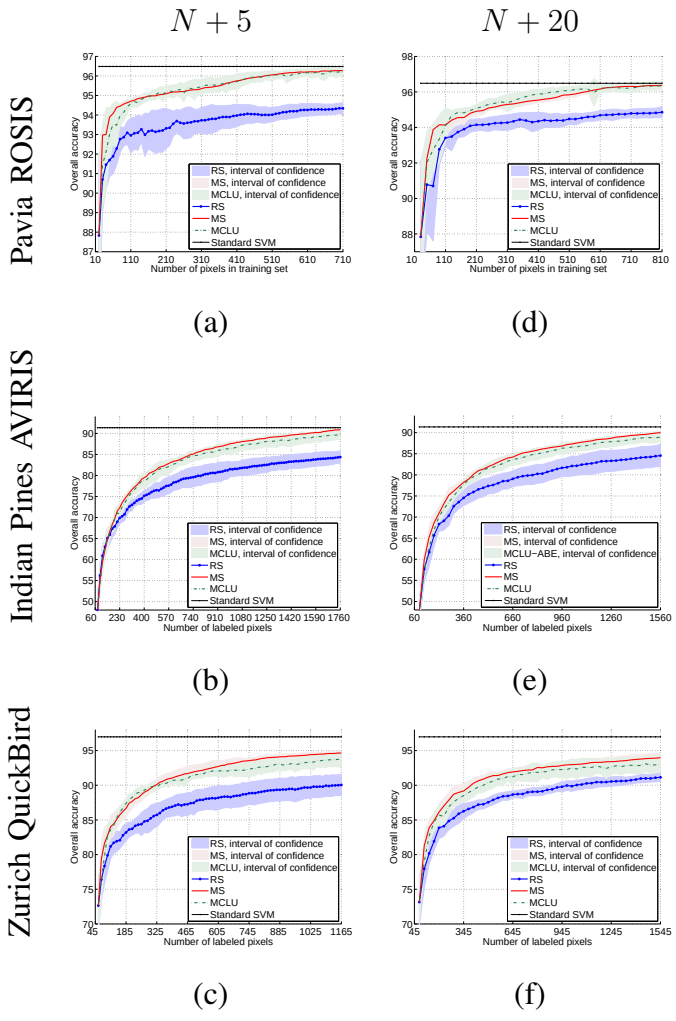



Fig. 5. Large margin active learning without diversity criterion. An example comparing MS and MCLU (RS = Random Sampling).

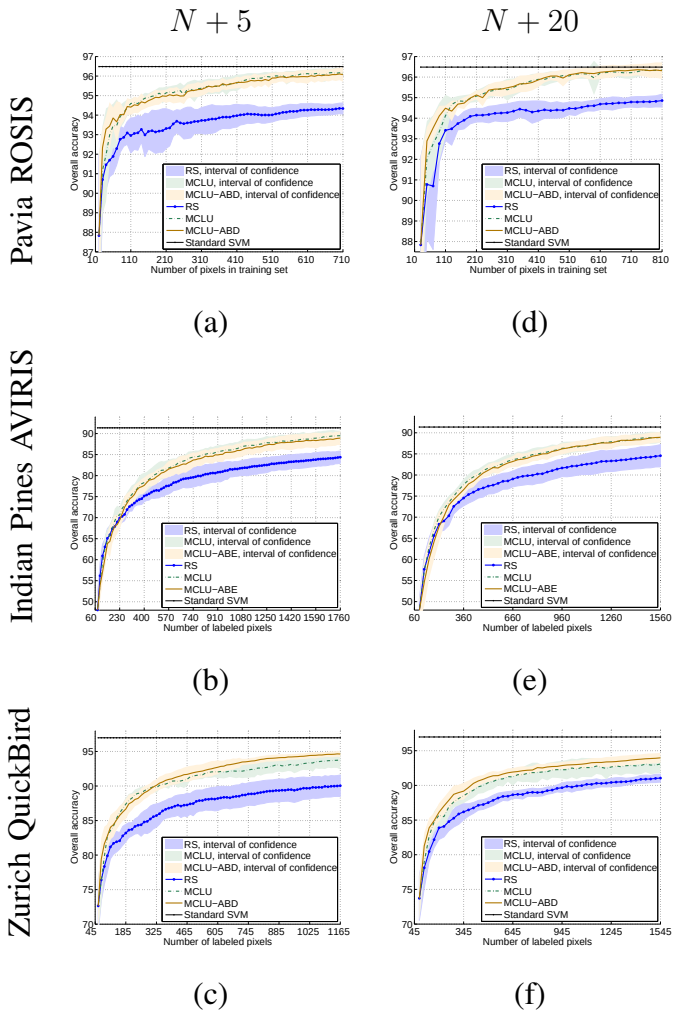


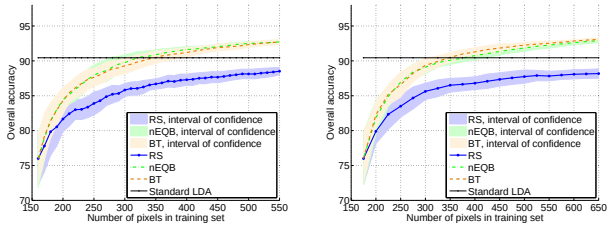

Fig. 6. Effect of diversity criteria on large margin active learning. An example comparing MCLU and MCLU-ABD (RS = Random Sampling).

a better generalization.

Summing up, when using methods other than large margin-based algorithms, performances of the heuristics are similar and the choice must be driven by the specific constraints of time and number of iteration allowed. We will come back to these issues in the next section.

## VIII. Discussion

Throughout the experiments presented, nearly all the algorithms compared showed fast convergence to the upper bounds represented by the Standard SVM/LDA. At convergence, all the heuristics outperformed the random selection of pixels. The interest of using multiclass




(a) $N + 5$ (b) $N + 20$

Fig. 7. Committee-based and posterior probability heuristics trained with LDA classifiers on the Pavia ROSIS image (RS = Random Sampling).

uncertainty or adding a criterion of diversity has been demonstrated by the experiments above. Table I summarizes the points raised in this paper and gives a general overview of the methods reviewed.

Large margin-based methods with diversity criterion seem the most appropriate when using SVM, while committee-based heuristic leave freedom to the user to exploit the model he is most confident with. Moreover, the user can build ensembles of classifiers exploiting the intrinsic advantages of specific classifiers for a given type of image. Weighted committees or boosting candidates are also possible (see Section III for some references). Probabilistic heuristics have the advantage of speed, but cannot always provide batches of samples and, if the classifier does not return such estimates naturally, must rely on further approximations of the posterior probabilities.

However, it would not be correct to base the choice of the heuristic in a model-based fashion only. The choice of the best heuristic is problem-oriented and depends on the needs of the user in terms of time, complexity and size of the batch to be provided at each iteration. This section draws some guidelines to select the most appropriate heuristic.

A first distinction could be done depending on the type of the images considered:

- when dealing with hyperspectral images, which are typically high dimensional, strategies taking direct advantage of the data structure should be preferred: typically, multi-view heuristics such as the AMD or the ECBD-ABD are particularly well-suited to this type of data. The first exploits cross-informations directly in the space of the spectral bands, while the second selects the samples according to spectral angles among the candidates.
- when the initial training set is very small, heuristics based on posterior probabilities should

be avoided, since such estimation strongly depends on the quality of the estimation of the class statistics (typically in the case of LDA). The same holds for committees-based on bagging, especially if the bags contain a small share of the original samples.

- when dealing with complex manifolds, in which redundancy can greatly affect the quality of the sampling and of the resulting training set, approaches based on modeling of the relationships among samples in the feature space can be strongly beneficial to select pixels reflecting such complexity. The use of kernel $k$-means in the MCLU-ECBD, in hMCS or the distance to support vectors in the cSV heuristic provide solutions in this sense.

A second distinction, more related to operational requirements, is based on the type of sampling to be performed by the user [39],

- when working by photointerpretation (typically in VHR imaging), sampling can be done on-screen directly by the user. This allows for large amounts of iterations and can thus be solved by small batches of samples. In this case complex heuristics favoring spectral diversity are to be preferred, since the complexity of the heuristics enforcing diversity strongly increases with the size of the batch considered.
- on the contrary, when sampling is to be done on the field (typically in hyperspectral or mid-resolution images), only a few iterations with large batches are allowed. In this case, all the heuristics seem to provide the same type of convergence and the user should prefer simple heuristics such as MCLU, BT or EQB, depending on the model used. In this case, the spatial location of samples seems to be much more important than the heuristic used: a pioneering work in this sense can be found in [54], where MS and BT are exploited with spatially adaptive cost and the sampling is optimized with respect to the spatial distance among the samples chosen.
- when sampling is done with moving sensors and the samples are acquired sequentially by an automatic device, batches of samples are not necessary. In this case models with small computational cost should be preferred, as they can update fast and almost instantly provide the next location to be sampled. In this case, BT and KL-max are most valuable.

## IX. Conclusion

In this paper we presented and compared several state of the art approaches to active learning for the classification of remote sensing images. A series of heuristics have been classified by

TABLE I

SUMMARY OF ACTIVE LEARNING ALGORITHMS (B : BINARY, M : MULTICLASS, $q$ : NUMBER OF CANDIDATES, $k$ : MEMBERS OF THE COMMITTEE OF LEARNERS, $S$ : BATCH, $SVs$ : SUPPORT VECTORS, ✓ : YES, × : NO).

| Family | Heuristic | Reference | Batches | Uncertainty | Classifier | Diversity | | Models to train |
|---|---|---|---|---|---|---|---|---|
| | | | | | | $S$ | $SV$s | |
| Committee | EQB | [21] | ✓ | M | All | × | × | $k$ models |
| | AMD | [28] | ✓ | M | All | × | × | $k$ models |
| Large margin | MS | [36] | ✓ | B | SVM | × | × | Single SVM |
| | MCLU | [39] | ✓ | M | SVM | × | × | Single SVM |
| | SSC | [41] | ✓ | B | SVM | × | × | 2 SVMs |
| | cSV | [21] | ✓ | B | SVM | ✓ | × | Single SVM + distances to support vectors |
| | MOA | [38] | ✓ | B | SVM | ✓ | × | Single SVM + distances to already selected pixels |
| | MCLU-ABD | [39] | ✓ | M | SVM | ✓ | × | Single SVM + distances to already selected pixels |
| | MCLU-ECBD | [39] | ✓ | M | SVM | ✓ | × | Single SVM + nonlinear clustering of candidates |
| | hMCS-i | [46] | ✓ | M | SVM | ✓ | ✓ | Single SVM + nonlinear clustering of candidates and SVs |
| Posterior probability | KL-max | [48] | × | M | $p(y\|\mathbf{x})$ | × | × | $(q-1)$ models |
| | BT | [51] | ✓ | M | $p(y\|\mathbf{x})$ | × | × | Single model |

their characteristics into four families. For each family, some heuristics have been detailed and then applied to three challenging remote sensing datasets for multispectral and hyperspectral classification. Advantages and drawbacks of each method have been analyzed in detail and recommendations for further improvement have been worded. However, this review is not exhaustive and the research in the field is far from being over: there is a healthy and rich research community developing new heuristics for active sampling that have been or will be presented in the remote sensing and signal processing community.

Active learning has a strong potential for remote sensing data processing. Efficient training sets are needed by the users, especially when dealing with large archives of digital images. New problems are being tackled with active learning algorithms, guaranteeing the renewal of the field.

Some recent examples can be found in the active selection of unlabeled pixels for semi-supervised classification [55], spatially adaptive heuristics [54] or the use of active learning algorithms for model adaptation across domains [49], [56]. Further steps for active learning methods are the inclusion of contextual information in the heuristics: so far, the heuristics proposed only take advantage of spectral criteria – or at most include contextual features in the data vector – but few heuristics directly consider positional information and/or textures. Another crucial issue is the robustness to noise: since they are based on the uncertainty of the pixels, current heuristics are useless for images related to high levels of noise such as SAR. This field remains, at present, totally unexplored.

## ACKNOWLEDGMENTS

The authors would like to acknowledge prof. Paolo Gamba (Univ. Pavia) who provided the Pavia dataset, as well as the authors in [53] for the Indian Pines data.

## REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Series in Statistics. Springer, New York, USA, 2nd edition, 2009.

[2] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. Annu. Workshop Comput. Learn. Theroy*, 1992, pp. 144–152.

[3] B. Schölkopf and A. J. Smola, *Learning with Kernels*, MIT press, Cambridge (MA), 2002.

[4] S. Haykin, *Neural Networks and Learning Machines*, Prentice Hall, 3rd edition, 2008.

[5] G. Camps-Valls and L. Bruzzone, *Kernel methods in Remote Sensing Image Processing*, Wiley and Sons, 2009.

[6] L. Bruzzone and L. Carlin, "A multilevel context-based system for classification of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 9, pp. 2587–2600, 2006.

[7] F. Pacifici, M. Chini, and W.J. Emery, "A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification," *Remote Sens. Environ.*, vol. 113, no. 6, pp. 1276–1292, 2009.

[8] D. Tuia, F. Pacifici, M. Kanevski, and W. J. Emery, "Classification of very high spatial resolution imagery using mathematical morphology and support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 47(11), pp. 3866 – 3879, 2009.

[9] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42(8), pp. 1778 – 1790, 2004.

[10] G. Camps-Valls, L. Gómez-Chova, J. Calpe-Maravilla, J. D. Martín-Guerrero, E. Soria-Olivas, L. Alonso-Chordá, and J. Moreno, "Robust support vector method for hyperspectral data classification and knowledge discovery," *IEEE Trans. Geosci. Remote Sens.*, vol. 42(7), pp. 1530 – 1542, 2004.

[11] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804 – 3814, 2008.

[12] G. M. Foody and A. Mathur, "Toward intelligent training of supervised image classifications: Directing training data acquisition for SVM classification," *Remote Sens. Env.*, vol. 93, pp. 107 – 117, 2004.

[13] G. M. Foody, A. Mathur, C. Sanchez-Hernandez, and D. S. Boyd, "Training set size requirements for the classification of a specific class," *Remote Sens. Env.*, vol. 104, pp. 1 – 14, 2006.

[14] G. M. Foody and A. Mathur, "The use of small training sets containing mixed pixels for accurate hard image classification: training on mixed spectral responses for classification by a SVM," *Remote Sens. Environ.*, vol. 103, no. 2, pp. 179–189, 2006.

[15] D. J. C. MacKay, "Information based objective functions for active data selection," *Neural Comp.*, vol. 4(4), pp. 590 – 604, 1992.

[16] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Mach. Learn.*, vol. 15(2), pp. 201–221, 1994.

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

[18] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Proc. 11th ICML*, 1994, pp. 148 – 156.

[19] F. Olsson, "A literature survey of active machine learning in the context of natural language processing," SICS Technical Report T2009:06, Swedish Institute of Computer Science,, 2009.

[20] B. Settles, "Active learning literature survey," Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2010.

[21] 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(7), pp. 2218 – 2232, 2009.

[22] D. Tuia, J. Muñoz-Marí, M. Kanevski, and G. Camps-Valls, "Cluster-based active learning for compact image classification," in *IEEE International Geoscience and Remote Sensing Symposium, IGARSS*, Hawaii, USA, 2010.

[23] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proc. Annu. Workshop Comput. Learn. Theroy*, 1992, pp. 287 – 294.

[24] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Mach. Learn.*, vol. 28, pp. 133 – 168, 1997.

[25] N. Abe and H. Mamitsuka, "Query learning strategies using boosting and bagging," in *Intl. Conf. Mach. Learn. ICML*, Madison, USA, 1998, pp. 1–9, Morgan Kaufmann.

[26] P. Melville and R. J. Mooney, "Diverse ensembles for active learning," in *Intl. Conf. Mach. Learn. ICML*, Banff, Canada, 2004, vol. 69 of *ACM International Conference Proceeding Series*, pp. 74–82, ACM Press.

[27] L. Copa, D. Tuia, M. Volpi, and M. Kanevski, "Unbiased query-by-bagging active learning for VHR image classification," in *Proceedings of the SPIE Remote Sensing Conference*, Toulouse, France, 2010.

[28] W. Di and M. Crawford, "Multi-view adaptive disagreement based active learning for hyperspectral image classification," in *IEEE International Geoscience and Remote Sensing Symposium, IGARSS*, Hawaii, USA, 2010.

[29] L. Breiman, "Bagging predictors," Tech. Rep. 421, University of California at Berkley, 1994.

[30] I. Muslea, "Active learning with multiple views," *Journal of Artificial Intelligence Research*, vol. 27, pp. 203–233, 2006.

[31] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45 – 66, 2001.

[32] C. Campbell, N. Cristianini, and A.J. Smola, "Query learning with large margin classifiers," in *Proc. 17th ICML*, 2000, pp. 111 – 118.

[33] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *Proc. 17th ICML*, 2000, pp. 839 – 846.

[34] S. Zomer, M. N. Sànchez, R. G. Brereton, and J.L. Pérez-Pavón, "Active learning support vector machines for optimal sample selection in classification," *J. Chemometrics*, vol. 18(6), pp. 294–305, 2004.

[35] S. Cheng and F. Y. Shih, "An improved incremental training algorithm for support vector machines using active query," *Patter Recogn.*, vol. 40, pp. 964 – 971, 2007.

[36] P. Mitra, B. Uma Shankar, and S.K. Pal, "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern Recogn. Lett.*, vol. 25, no. 9, pp. 1067–1074, 2004.

[37] L. Bruzzone and C. Persello, "Active learning for classification of remote sensing images," in *IEEE International Geoscience and Remote Sensing Symposium, IGARSS*, Cape Town, South Africa, 2009.

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

[39] 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.*, in press.

[40] A. Vlachos, "A stopping criterion for active learning," *Comput. Speech Lang.*, vol. 22, no. 3, pp. 295–312, 2008.

[41] E. Pasolli and F. Melgani, "Model-based active learning for SVM classification of remote sensing images," in *IEEE International Geoscience and Remote Sensing Symposium, IGARSS*, Hawaii, USA, 2010.

[42] Klaus Brinker, "Incorporating diversity in active learning with support vector machines," in *Proc. 20th ICML*, 2003.

[43] M. Girolami, "Mercer kernel-based clustering in feature space," *IEEE Trans. Neural Net.*, vol. 13(3), pp. 780 – 784, 2002.

[44] I. Dhillon, Y. Guan, and B. Kulis, "A unified view of kernel k-means, spectral clustering and graph cuts," Tech. Rep. UTCS Technical Report No. TR-04-25, University of Texas, Austin, Departement of Computer Science, 2005.

[45] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[46] M. Volpi, D. Tuia, and M. Kanevski, "Cluster-based active learning for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, submitted.

[47] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," in *Intl. Conf. Mach. Learn. ICML*, Williamstown (MA), USA, 2001, pp. 441–448, Morgan Kaufmann.

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

[49] G. Jun and J. Ghosh, "An efficient active learning algorithm with knowledge transfer for hyperspectral data analysis," in *IEEE Geosci. Remote Sens. Symp. IGARSS*, 2008.

[50] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*. MIT press, 1999.

[51] T. Luo, K. Kramer, D. B. Golgof, 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.

[52] G. Licciardi, F. Pacifici, D. Tuia, S. Prasad, T. West, F. Giacco, J. Inglada, E. Christophe, J. Chanussot, and P. Gamba,

"Decision fusion for the classification of hyperspectral data: Outcome of the 2008 GRS-S data fusion contest," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3857–3865, 2009.

[53] Q. Jackson and D. Landgrebe, "An adaptive classifier design for high-dimensional data analysis with a limited training data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 12, pp. 2664–2679, 2001.

[54] A. Liu, G. Jun, and J. Ghosh, "Active learning with spatially sensitive labeling costs," in *NIPS Workshop on Cost-sensitive Learning*, 2008.

[55] Jun Li, J.M. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4085 –4098, 2010.

[56] D. Tuia, E. Pasolli, and W. J. Emery, "Dataset shift adaptation with active queries," in *URBAN /URS Joint Event*, Munich, Germany, 2011.