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, and Jordi Muñoz-Marí

Josu Maiora

Overview

- Defining an efficient **training set** →Fundamental phase for classification
- 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**
- This paper reviews and tests the main families of active learning algorithms:
    1. committee,
    2. large margin,
    3. posterior probability-based

1. CONCEPTS AND DEFINITIONS

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^ϵ.
  for each candidate in U^ϵ **do**
    Evaluate a user-defined *heuristic*
  end for
  Rank the candidates in U^ϵ 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.

2. COMMITTEE-BASED ACTIVE LEARNING

The first family of active learning methods quantifies the uncertainty of a pixel by considering a committee of learners.

1. *Normalized Entropy Query-by-Bagging*

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 labels of the candidates.

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

where

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

where

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

2. COMMITTEE-BASED ACTIVE LEARNING

2. *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.*

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

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

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

where

$$p^{\mathrm{MV}}\left(y_i^* = \omega \mid \mathbf{x}_i^v\right) = \frac{\sum_{v=1}^{V} W^{\epsilon-1}(v,\omega)\delta(y^*_{i,v},\omega)}{\sum_{v=1}^{V}\sum_{j=1}^{N_i} W^{\epsilon-1}(v,\omega)}$$

3. LARGE-MARGIN-BASED ACTIVE LEARNING

The second family of methods is specific to margin-based classifiers (SVM)

The distance of a sample $\boldsymbol{x}_i$ from the SVM hyperplane is given by

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

1. *Margin Sampling (MS)*

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

2. *Multiclass Level Uncertainty (MCLU)*

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

where

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

3. LARGE-MARGIN-BASED ACTIVE LEARNING

3. *Significance Space Construction (SSC)*

The support vector coefficients are used to convert the multiclass classification problem into a binary support vector detection problem. This second classifier predicts which pixels are likely to become support vectors:

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

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

4. On the Need for a Diversity Criterion

- The heuristic, called “most ambiguous and orthogonal” (MAO) is iterative: starting from the samples selected by MS, , this heuristic iteratively chooses the samples minimizing the highest values between the candidates list and the samples already included in the batch .

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

- the MAO criterion is combined with the MCLU uncertainty estimation in the “multiclass level uncertainty—angle based diversity” (MCLU-ABD) heuristic.

$$\hat{\mathbf{x}}^{\mathrm{MCLU-ABD}} = \arg \min_{\mathbf{x}_i \in U^{\mathrm{MCLU}}} \left\{ \lambda f(\mathbf{x}_i)^{\mathrm{MC}} + (1-\lambda) \max_{\mathbf{x}_j \in S} \frac{K(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{K(\mathbf{x}_i, \mathbf{x}_i) K(\mathbf{x}_j, \mathbf{x}_j)}} \right\} \quad (12)$$

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

4. *On the Need for a Diversity Criterion*

- *Constraining the MS solution to pixels associated to different closest support*

$$\hat{\mathbf{x}}^{\mathrm{cSV}} = \arg \min_{\mathbf{x}_i \in U^{\mathrm{MS}}} \{ |f(\mathbf{x}_i, \omega)| \,|\, cSV_i \notin cSV_\theta \} \qquad (13)$$

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*

$$\hat{\mathbf{x}}^{\mathrm{MCLU-ECBD}} = \arg \min_{\mathbf{x}_i \in c_m} \{ f(\mathbf{x}_i)^{\mathrm{MC}} \},$$
$$m = [1, \ldots q], \quad \mathbf{x}_i \in U^{\mathrm{MCLU}} \qquad (14)$$

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

4. POSTERIOR PROBABILITY BASED ACTIVE LEARNING

- The third class of methods uses the estimation of posterior probabilities of class membership (i.e.,*P(y/x)*) to rank the candidates.
  - *KL-Max*

    The first idea is to sample the pixels whose inclusion in the training set

$$\hat{\mathbf{x}}^{\mathrm{KL-max}} = \arg\max_{\mathbf{x}_i \in U} \left\{ \sum_{\omega \in N} \frac{1}{(u-1)} \times \mathrm{KL}\left(p^{+}(\omega \mid \mathbf{x}) \middle|\middle| p(\omega \mid \mathbf{x}) \right) p(y_i^* = \omega \mid \mathbf{x}_i) \right\} \quad (16)$$

where the condition $n_{c_m}^{bSV} = 0$ ensures that the cluster queried does not contain any bounded support vector sampled at the previous iteration

4. POSTERIOR PROBABILITY BASED ACTIVE LEARNING

- *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 for each candidate .

In this case, the per-class posterior probability is assessed fitting a sigmoid function to the SVM decision function [50]:

$$p(y_i^* = \omega \mid \mathbf{x}_i) = \frac{1}{1 + e^{(Af(\mathbf{x}_i, \omega) + B)}} \tag{18}$$

$$\hat{\mathbf{x}}^{\mathrm{BT}} = \arg \min_{\mathbf{x}_i \in U} \left\{ \max_{\omega \in N} \{p(y_i^* = \omega \mid \mathbf{x}_i)\} - \max_{\omega \in N \setminus \omega^+} \{p(y_i^* = \omega \mid \mathbf{x}_i)\} \right\} \tag{19}$$

4. DATASETS

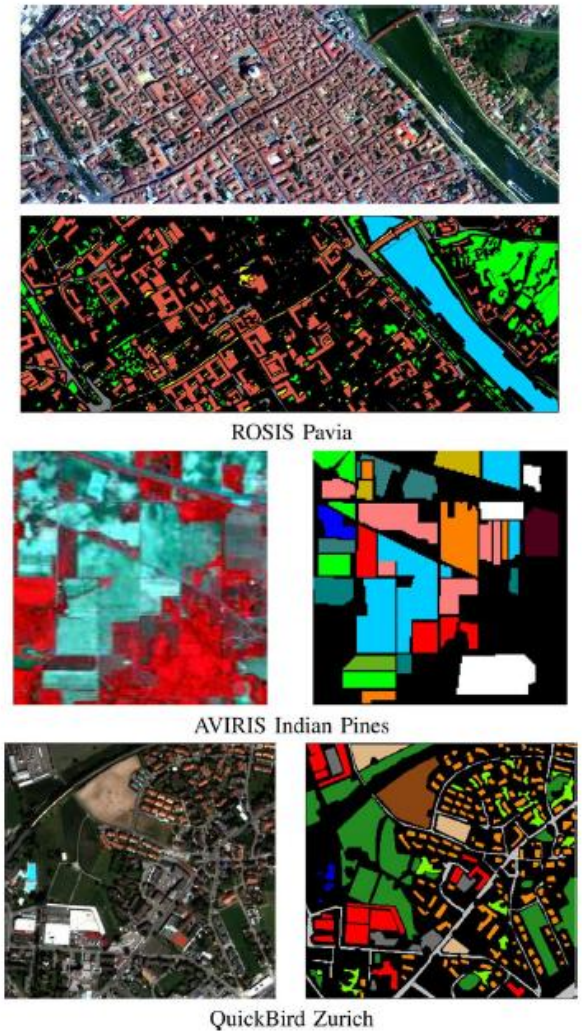


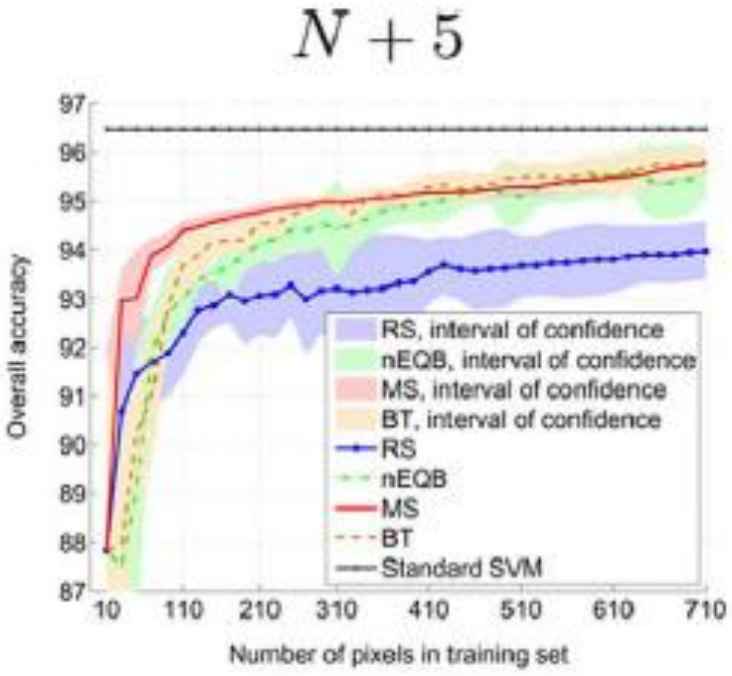
Fig. 2. Images considered in the experiments: (top) ROSIS image of the city of Pavia, Italy (bands $[56 - 31 - 6]$ and corresponding ground survey); (middle) AVIRIS Indian Pines hyperspectral data (bands $[40 - 30 - 20]$ and corresponding ground survey); (bottom) QuickBird multispectral image of a suburb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

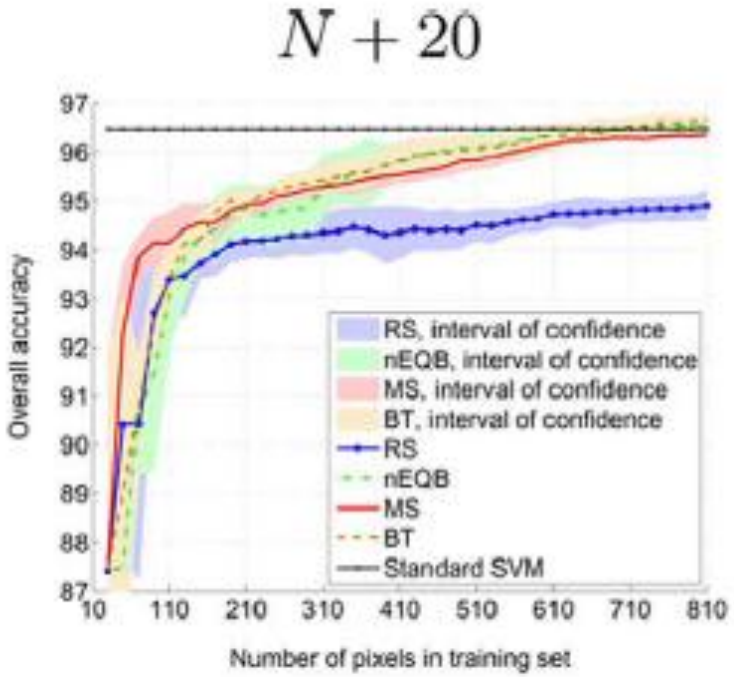
5. EXPERIMENTAL SETUP

- 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 **five-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.

6. NUMERICAL RESULTS

Pavia ROSIS





5. CONCLUSION

- A series of heuristics have been classified by their characteristics into three families.
- Active learning has a strong potential for remote sensing data processing.
- Some recent examples can be found in the active selection of unlabeled pixels for semi-supervised classification.