CLUSTER-BASED ACTIVE LEARNING FOR COMPACT IMAGE CLASSIFICATION

Devis Tuia [1], *Mikhail Kanevski* [1], *Jordi Muñoz Marí* [2] *and Gustavo Camps-Valls* [2]

[1] Institute of Geomatics and Analysis of Risk, University of Lausanne, Switzerland.
[2] Image Processing Laboratory (IPL), University of València, Spain.

1. INTRODUCTION

Active learning deals with developing methods that select examples that may express data characteristics in a compact way [1]. In recent years, increasing attention has been paid to active learning methods in the remote sensing community. Strategies for intelligent sampling have been proposed with model-based heuristics aiming at the efficient search of the best pixels to be labeled to increase model's performance [2–4]. However, all these strategies rely on i) an initial training set obtained randomly from all the unlabeled pixels and ii) assume that all the important regions of the feature space have been covered by such set and concentrate on uncertain regions of the current classifier to reduce the hypothesis space (uncertainty-based active learning [5,6]). However, this last assumption is really strong and can make the algorithm fail to converge in some particular settings. A typical situation for non-convergence is met when the initial distribution of the unlabeled pixels is biased: in that case, the initial description of the data will ignore small regions, and these regions will be ignored by uncertainty-based heuristics. In this paper, we propose the use of the cluster structure of data to guide the active learning strategy through an initial set thus ensuring convergence. The aim is to optimize the generation of the initial training set in order to cover all the interesting regions of the feature space. In [7] a pre-clustering of the data was proposed for the initial samples. Yet interesting, this strategy only optimizes the covering of the space, but the initial sampling still remains random. As a solution, in [8] a semi-supervised strategy is proposed: using a hierarchical clustering of the data, the best cluster structure for the sampling is found by active queries. Using the cluster structure encoded in the clustering result, the queries are gradually directed toward the regions of the cluster structure showing mixed labeling. In this paper, we explore this last strategy for remote sensing image classification.

2. ACTIVE HIERARCHICAL SAMPLING

In this section we motivate and revise the main idea under the active algorithm proposed.

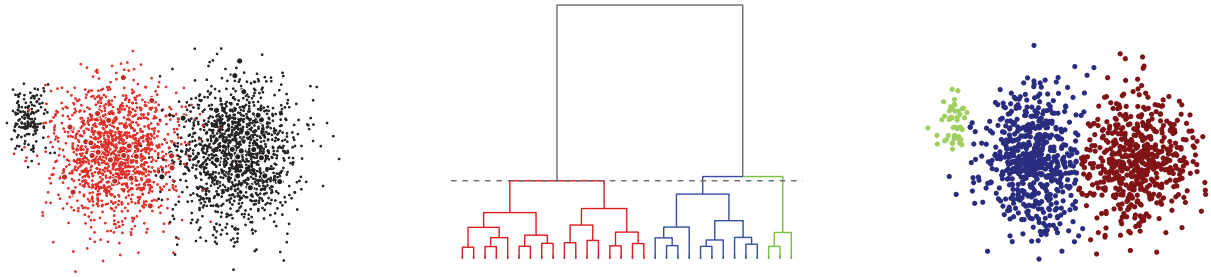
2.1. Why and when data structure counts?

Consider a two-classes problem with clustered distributions and where the data distribution is strongly unbalanced, as shown in Fig. 1. Since the initial training samples (filled points in Fig. 1a) do not consider it, the left black cluster will not be sampled. This situation is quite common because with random sampling the probability of picking a pixel is proportional to the size of the cluster, which is definitely a bad choice.

An uncertainty-based strategy will ignore the left cluster and will not converge to an optimal solution. This is because such an approach will direct the sampling into the region where the decision function lies. Nonetheless, a hierarchical clustering algorithm can detect such a structure (Fig. 1b) and, using the resulting clustering in the input space, the sampling can be directed to that region (Fig. 1c).

The open question is: given a hierarchical structure (or a *tree*), how to prune it to detect the correct data structure? On one hand, if the tree is too general, the subregions will not be sampled quicker than the random strategy and, on the other hand, if the hierarchical tree is too much pruned, the strategy will not be efficient, because several queries will be needed.

This paper is partially funded by the Swiss National Foundation, under grants no. PBLAP2-127713 and 200021-126505.



(a) (b) (c)

Fig. 1. Toy example for a biased distribution of the unlabeled pixels. (a) Distribution of the candidates in the input space (2-classes problem); (b) hierarchical structure related to the clustering of the optimal search regions in (c).

2.2. Integrating data structure in random queries

To answer this question, the (unlabeled) clustering tree must be queried iteratively and with the labels found, the algorithm will decide if it is necessary to prune the tree further in order to clarify the labeling of the clusters (hereafter the *nodes*).

Starting with a given pruning V, we can estimate the probability of a node v of size n_v to belong to a class l by using the queried leaves (pixels). The quality of such an assessment at time t depends of the number of leaves consulted and can be assessed using generalization error bounds [8]:

$$[p_{v,l}^{LB}(t), p_{v,l}^{UB}(t)] = [\max(p_{v,l}(t) - \Delta_{v,l}(t), 0),\ \min(p_{v,l}(t) + \Delta_{v,l}(t), 1)]\,, \tag{1}$$

where $\Delta_{v,l}(t) \approx \frac{1}{n(t)} + \sqrt{\frac{p_{v,l}(t)(1-p_{v,l}(t))}{n_v(t)}}$. A given label is defined to be *admissible* for a given pruning if $p_{v,l}^{LB}(t) > 2p_{v,l'}^{UB} - 1$, $\forall l' \neq l$. In other words, l is an admissible label for node v if it gives at most twice as much error as any other label [8]. Once all the admissible couples $\{v, l\}$ have been found, the cost of the different pruning is evaluated as $1 - p_{v,l}(t)$. If pruning a node into its two sub-nodes decreases the global cost, the tree is modified accordingly.

2.3. Active strategies

So far, we have a clear strategy to divide the tree when the labels observed are mixed: if the division decreases the cost, i.e. if the sub-clusters are more pure, the tree is pruned. The active queries play a role in the selection of the node to sample. A first, trivial, strategy, would be to sample according to the size of the clusters. This strategy is close to random sampling, and will be used as a comparison in the following experiments. More intelligent strategies can be used by accounting for the bounds of Eq. (1). The first approach, called $CS - s1$, selects the node to be sampled proportionally to its size and the certainty of its label, assessed by the lower bound on the winning class [8]:

$$v = \max_{\forall v \in V} (1 - p_{v,l}^{LB}(t))\, n_v(t) \tag{2}$$

This strategy takes advantage of large, pure, clusters: as soon as they are considered pure enough, the sampling is directed elsewhere. A second strategy, called $CS - s2$ consists of excluding the size term of Eq. (2): in this case, as soon as a cluster is considered uncertain, it is divided and sampled until the purity of the sub-tree is considered satisfactory.

3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1. Data and setup

The experiments have been performed on a multispectral image acquired over Zurich, Switzerland. The image was taken by the QuickBird sensor and has been pan sharpened to a spatial resolution of 0.6 m. Four spectral channels are used for classification $\{B, G, R, NIR\}$. The image shows a typical urban neighborhood, with four predominant land use classes: built-up, roads, vegetation and water. Due to the particular illumination conditions, a fifth class, accounting for shadows, has been labeled.

The proposed CS active method is unsupervised. Therefore, it starts with no labeled information. The model has been compared with a random sampling starting with no labeled pixels. For the CS model, the C++ implementation available at

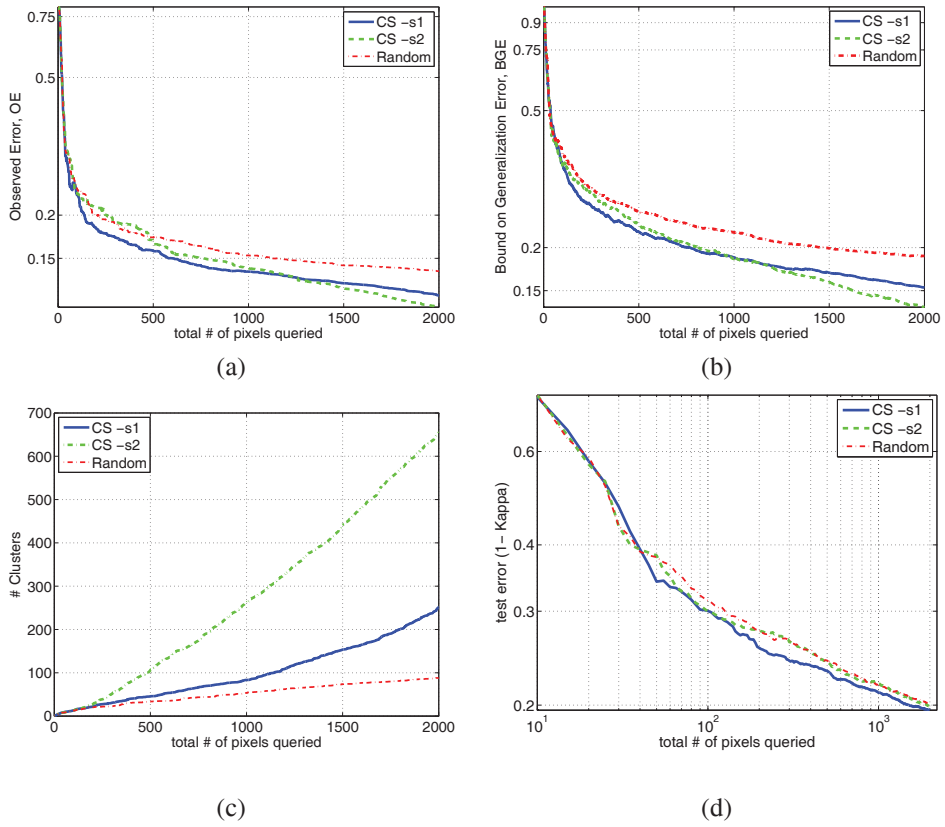


Fig. 2. Results of the cluster-based active learning. (a) Observed error on the pool of candidates; (b) bound on generalization error (BGE); (c) number of clusters pruned; (d) Test performance (1 - Kappa) obtained by KNN ($k = 5$) using the training sets obtained by the different strategies.

`http://cseweb.ucsd.edu/~djhsu` of [8] has been used. The tree structure has been computed using standard linkage with Euclidean distance and Ward's aggregation rule (Matlab function `linkage.m`).

3.2. Results and discussion

Figure 2 illustrates the results for the CS strategies proposed. First, a transductive setting is considered: in Figs. 2(a) and 2(b) only the membership of the pixels on the tree is predicted. The observed error (computed using the overall accuracy OA as $1 - OA/100$), as well as the bound on the generalization error $\frac{1}{n}\sum_{i=1}^{n}(1 - p_{i,l}^{LB}(t))$ are considered respectively. Figure 2(c) shows the number of cluster for each strategy. Note this transductive setting yields the expected result, as the whole image should be clustered, and thus *no model training is required in this setting*. Then, we used a k-NN classifier with the k parameter fixed to 5 to predict an unknown test dataset (Fig. 2(d)). This second setting has been considered for computational reasons due to that the pool of candidates –and consequently the tree– only exploits 10,000 pixels. At first glance, both active strategies show good performances with respect to random sampling in the transductive setting. The $CS - s1$ shows faster convergence from the first iterations, while the $CS - s2$ strategy takes more time to find the optimal solution and converges to lower classification errors. This is related to the difference in the sampling strategies: $CS - s1$ keeps the number of clusters as low as possible by pruning only when a large cluster's class becomes dubious and directing the sampling elsewhere after pruning. On the contrary, the $CS - s2$ strategy divides the uncertain cluster several times until a stable solution is found, see Fig. 2c. Regarding the generalization on new sets, Fig. 2d, strategy $CS - s1$ provides the best results because it avoids sampling outliers as $CS - s2$ does.

This difference among the active strategies proposed can be observed in Fig. 3. The $CS - s1$ strategy (Fig. 3a) gives priority to the difficult classes and samples more the mixed clusters containing pixels of buildings, roads and shadows. The classes vegetation and water is quickly abandoned, since its average reflectance makes the cluster assignment quite obvious. Looking at the $CS - s2$ results (Fig. 3b), a different trend is observed: the sampling is done by waves, first between the classes

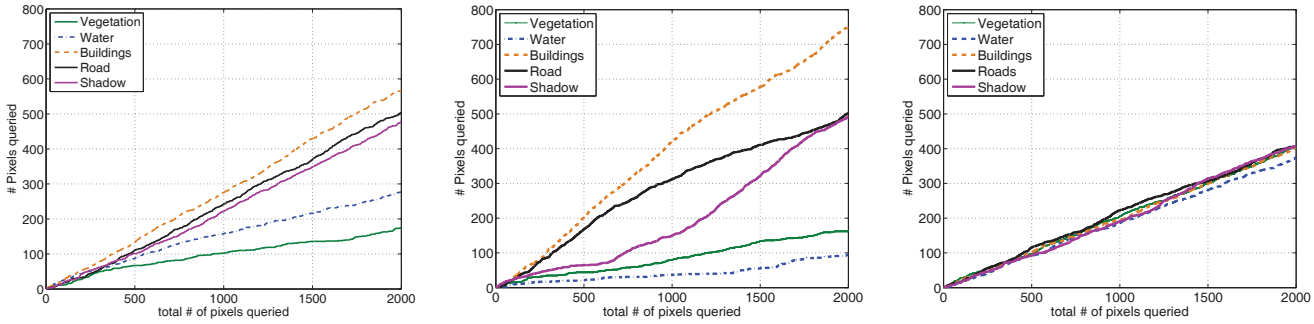


Fig. 3. Number of pixels per class sampled by the (a) $CS-s1$, (b) $CS-s2$ and (c) random strategies (for a single experiment).

building and road, then between buildings and shadows. By sequential pruning, the detail of the class boundary between single classes is refined. As a comparison, Fig. 3c shows the results obtained by random selection, for which all classes are sampled uniformly.

4. CONCLUSIONS

In this paper, a cluster-based strategy for active querying is presented. The method is unsupervised and does not require any model training. The proposed method exploits cluster structure of data and samples the nodes of a hierarchical structure by assessing the probability bounds on the class membership of the sub-clusters. Ongoing work is concerned with the study of clustering methods beyond linear linkage to better describe the structure of data in the case of complex data manifolds. More experimental results and theoretical details will be given at the time of the conference.

5. REFERENCES

[1] D. Cohn, L. Atlas, and Ladner R., "Improving generalization with active learning," *Mach. Learn.*, vol. 15, no. 2, pp. 201–221, 1994.

[2] 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.

[3] 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.

[4] D. Tuia, F. Ratle, F. Pacifici, M. Kanevski, and W.J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2218–2232, 2009.

[5] G. Schohn and D. Cohn, "Less is more: active learning with support vectors machines," in *Intl. Conf. Mach. Learn. ICML*, Stanford, USA, 2000, pp. 839–846, Morgan Kaufmann.

[6] S. Tong and D. Koller, "Support vector machines active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, 2002.

[7] S. Cheng and F.Y. Shih, "An improved incremental training algorithm for support vector machines using active query," *Pattern Recogn.*, vol. 40, no. 3, pp. 964–971, 2007.

[8] S. Dasgupta and D. Hsu, "Hierarchical sampling for active learning," in *Intl. Conf. Mach. Learn. ICML*, Helsinki, Finland, 2008, vol. 307 of *ACM International Conference Proceeding Series*, pp. 208–215, ACM Press.