# RECENT TRENDS IN CLASSIFICATION OF REMOTE SENSING DATA: ACTIVE AND SEMISUPERVISED MACHINE LEARNING PARADIGMS

*Lorenzo Bruzzone and Claudio Persello*

Department of Information Engineering and Computer Science, University of Trento
Via Sommarive, 14 I-38123, Povo, Trento, Italy, Fax: +39-0461-882093,
e-mail: lorenzo.bruzzone@ing.unitn.it, claudio.persello@disi.unitn.it

**Preference**: Overview presentation for the special session on “Machine Learning and Remote Sensing”

## 1. INTRODUCTION

Machine learning techniques have been widely used in the past decades for the analysis and classification of remote sensing (RS) data. In the early nineties, a large attention has been devoted to neural networks paradigms, which were widely used for RS image classification. Many different neural models (e.g., multilayer perceptrons, radial basis function, etc.) have been proposed, investigated and used for classification of different types of RS data (e.g., multispectral, hyperspectral, SAR images, etc.). In this framework, in the recent years, kernel methods [and especially support vector machines (SVMs)] have gained increasing attention in the RS community, and nowadays are becoming the state-of-the-art techniques replacing more traditional neural classifiers.

The development of the machine learning methodologies has been followed by a parallel development of novel concepts and paradigms that have gained some relevance also in the RS community. Since few years ago, automatic classification of RS images was typically performed by using supervised classification techniques, which require the availability of labeled samples for training the supervised algorithm. However, the collection of labeled samples is usually time consuming and costly and can be derived by: i) *in situ* ground truth surveys, ii) analysis of reliable portions of reference maps (when available), or iii) image photointerpretation. The amount and the quality of the available training samples are crucial for obtaining accurate classification maps. Nonetheless, in many real world problems the available training samples are not enough for an adequate learning of the classifier. The alternative unsupervised approach, which is implemented according to clustering methods, has been widely investigated in the past, but usually it is not reliable for obtaining an accurate categorization of the large number of information classes often present in RS classification problems.

In order to enrich the information given as input to the training phase of a supervised learning algorithm (and to improve classification accuracy), in the recent few years semisupervised approaches have been proposed, which jointly exploit labeled and unlabeled samples in the training of the classifier. Semisupervised approaches based on SVMs have been successfully applied to the classification of multispectral and hyperspectral data, where the ratio between the number of training samples and the number of available spectral channels is small [1].

However, an alternative and conceptually different approach for improving the statistic in the learning of a classifier is to iteratively expand the original training set according to an interactive process that involves a supervisor. This approach is known as active learning (AL) [2], and although still marginally considered in the RS community [3]-[5], can result very effective in different application domains. In AL: i) the learning process repeatedly queries available unlabeled samples to select the ones that are expected to be the most informative for an effective learning of the classifier, ii) the supervisor (e.g., the user) labels the selected samples interacting with the system, and iii) the learner update the classification rule by retraining with the updated training set. Therefore, the unnecessary and redundant labeling of non informative samples is avoided, greatly reducing the labeling cost and time. Moreover, AL allows one to reduce the computational complexity of the training phase. The important difference between the semisupervised approach and the AL is that semisupervised techniques usually automatically iterate by labeling and incorporating originally unlabeled samples in the training process (without requiring any additional effort from the user), whereas the AL approach requires the interaction between the system and the user, which is guided by the system to annotate unlabeled samples that are selected by the query function as the most informative for a complete modeling of the classification problem.

Despite the aforementioned difference, AL and semisupervised learning have some common theoretical aspects and properties, and plays around similar concepts exploited from two different perspective. In this paper we present a theoretical analysis on the two considered approaches in relation to the application to classification of remotely sensed images. Moreover, we present experimental comparisons that identify the advantages and disadvantages of AL and semisupervised methods, and also try to point out the boundary condition on the applicability of these methods both in terms of available labeled samples and reliability of classification results. Limits and potentials of both approaches are critically analyzed in the light of possible applications to different RS scenarios characterized by different kinds of data and classification problems.

## 2. PROBLEM FORMULATION AND METHODOLOGY

In order to formalize the considered problem, we can model an active learner as a quintuple ($C$, $Q$, $S$, $L$, $U$) [2]. $C$ is a supervised classifier, which is trained on the labeled training set $L$. $Q$ is a query function used to select the most informative unlabeled samples from an unlabeled sample pool $U$. $S$ is a supervisor who can assign the true class label to any unlabeled sample (e.g., a user that can derive the land cover type of the area on the ground associated to the selected pattern). After the initialization stage, the query function $Q$ is used to select a set of samples from the pool $U$ and the supervisor $S$ assigns them the true class label. Then, these new labeled samples are included into $L$ and the classifier $C$ is retrained using the updated training set. The closed loop of querying and retraining continues for some predefined iterations or until a stop criterion is satisfied.

It is worth noting that a (simple) semisupervised learner can be modeled with the same quintuple as for an active learner, by considering that the supervisor $S$ coincide with the classifier $C$. In this case the unlabeled samples selected by the query function are labeled directly by the classifier *C* (these samples are called semilabeled) or by implicit labeling strategies related to the classification procedure (e.g. exploitation of the cluster assumption), without requiring an external supervisor *S*. Nevertheless, the classifier *C* is not completely reliable (like the external supervisor *S*), and the query function *Q* of standard semisupervised techniques typically selects the most certain samples among the informative ones, instead of the most informative. Semisupervised approaches should operate a more “prudent” selection of unlabeled samples with respect to AL in order not to introduce semilabeled samples with the wrong semilabel and thus decreasing the performance of the classification system. This risk is not taken into account in AL, as *S* is considered able to assign the correct label to any unknown pattern. Moreover, many semisupervised approaches (e.g., transductive SVM [1]) employ a criterion to discard not reliable semilabeled samples, i.e., samples that change their semilabel in successive iterations. For the aforementioned reasons, a semisupervised approach usually requires more iterations (and samples) than an active learner in order to reach the convergence.

Both the AL and semisupervised processes require a first stage for the initialization of the training set. The main problem at the first stage is how many samples an initial training set should contain, which has an impact on the cost and the performances of the classification system. The cost associated to the definition of the initial training set mainly depends on the strategy adopted for the labeling process, i.e., photointerpretation, reference map analysis, or ground survey. The impact of the number (and quality) of the initial training samples on the performances of the learning system is very different for an active learner and for semisupervised techniques. In the AL case, too few samples may result in an incorrect query function at the first iterations, resulting in a slow increase of the accuracies in the first iterations. This usually affects the number of iterations, but not the convergence capability of the learning algorithm. On the contrary, the initial training set is much more critical for a semisupervised technique. Few and/or not representative samples may not allow a reasonable initial learning of the classifier, thus affecting, the convergence capability of the semisupervised learning process. Moreover, the semisupervised approach relies on the cluster assumption, i.e., samples of different classes belong to different clusters in the feature space. Nevertheless, under the appropriate hypothesis, a semisupervised approach allows one to increase the classification accuracy, without requiring any additional effort from an external supervisor *S*, i.e., the user. Given the aforementioned observations, here we present and discuss also an hybrid approach that combines the AL and the semisupervised paradigms in order to integrate the advantages of both approaches. In such an iterative learning procedure, some iterations are carried out according to an AL procedure (the query function selects the most informative samples to be labeled by an external supervisor *S*) and some iterations according to a semisupervised procedure (the query function select informative samples with good confidence on their label) without requiring the an external supervisor *S*.

## 3. EXPERIMENTAL RESULTS AND CONCLUSION

We applied both AL and semisupervised techniques based on SVMs to binary classification problems. We considered both toy examples and real RS problems in order to study and asses in detail the behavior of AL and semisupervised methods in different conditions. The toy examples were designed for emphasizing the understanding of the boundary conditions for a correct convergence of the learning procedure in semisupervised methods with respect to the number of available labeled samples and their capability to model the considered classification problem. Moreover, we compared semisupervised learning with different AL methods for identifying the tradeoff among computational time, reliability and labeling cost, which drives the choice between semisupervised and AL strategies. The analysis of the obtained results allows one to understand the limits of the semisupervised approach, i.e., in which conditions a semisupervised method is appropriate to extract information from unlabeled samples (given the available training set), and when an external supervisor is necessary to annotate (with the true label) the most critical samples thus defining an AL procedure. This analysis is very important for a proper understanding of the two different approaches and for deriving important indications to drive the user to choose one of the two strategies in real applications.

In the study, we also assessed the effectiveness of hybrid solutions, which integrate AL and semisupervised methods for taking advantage from their properties limiting the related drawbacks, i.e., the high cost associated with the annotation of a large number of unlabelled samples for AL, and the long computational time and the possible unreliability for semisupervised learning. Results point out the interesting indications for an effective use of the aforementioned hybrid strategy.

## REFERENCES

[1] L. Bruzzone, M. Chi, M. Marconcini, “A Novel Transductive SVM for the Semisupervised Classification of Remote-Sensing Images”, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 44, No. 11, pp. 3363-3373, 2006.

[2] M. Li and I. Sethi, “Confidence-Based active learning”, *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, No. 8,* pp. 1251-1261, 2006.

[3] P. Mitra, B. U. Shankar, and S. K. Pal, “Segmentation of multispectral remote sensing images using active support vector machines,” *Pattern Recognition Letters*, vol. 25, no. 9, pp. 1067–1074, Jul. 2004.

[4] S. Rajan, J. Ghosh, and M. M. Crawford, “An active learning approach to hyperspectral data classification,” *IEEE Transaction on Geoscience and Remote Sensing*, vol. 46, no. 4, pp. 1231-1242, Apr. 2008.

[5] D. Tuia, F. Ratle, F. Pacifici, M. Kanevski, and W. J. Emery, “Active Learning methods for remote sensing image classification,” *IEEE Transaction on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218 -2232, Jul. 2009.