ACTIVE AND SEMISUPERVISED LEARNING TECHNIQUES FOR CLASSIFICATION OF REMOTE SENSING IMAGES

Lorenzo Bruzzone and Claudio Persello

Department of Information Engineering and Computer Science, University of Trento
Via Sommarive, 14 I-38100, Povo, Trento, Italy, Fax: +39-0461-882093,
e-mail: lorenzo.bruzzone@ing.unitn.it, claudio.persello@disi.unitn.it

Preference: Oral presentation

1. INTRODUCTION

Automatic classification of remote sensing (RS) images is typically performed by using supervised classification techniques, which require the availability of labeled samples for training the supervised algorithm. 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 photo-interpretation. The amount and the quality of the available training samples is crucial for obtaining accurate classification maps. However, in many real world problems the available training samples are not enough for an adequate learning of the classifier. In order to enrich the information given as input to the learning algorithm (and to improve classification accuracy) semisupervised approaches can be adopted to jointly exploit labeled and unlabeled samples in the training of the classifier. Semisupervised approaches based on Support Vector Machines (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 expands the original training set according to an interactive process that involves a supervisor. This approach is known in the machine learning community as active learning [2], and although marginally considered in the remote sensing community [3], can result very effective in different application domains. In active learning: 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, active learning allows one to reduce the computational complexity of the training phase. The important difference between the semisupervised approach and the active learning is that semisupervised techniques automatically iterate by labeling and incorporating originally unlabeled samples in the training process (without requiring any additional effort from the user), whereas the active learning 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, active learning and semisupervised learning have some common theoretical aspects and properties. In this paper we present a theoretical and experimental analysis on the two considered approaches, studying their application to classification of remotely sensed images. Moreover, we propose two classification algorithms based on SVM for the analysis of RS images that implement two novel active and semisupervised learning strategies. Limits and potentials of both approaches are compared and critically analyzed in the light of possible applications.

2. PROBLEM FORMULATION AND METHODOLOGY

In order to formalize our problem, we can model an active learner as a quintuple (C, Q, S, L, U). 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). 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, without requiring an external supervisor S. Nevertheless, the classifier C is not completely reliable (like the external supervisor), and the query function Q of standard semisupervised techniques typically selects the most certain samples among the informative ones (instead of the most informative).

An active learning process requires a first stage for the initialization of the training set, and a two-steps closed-loop stage of query of unlabeled samples and retraining of the supervised learner. The main problem at the first stage is how many samples an initial training set should contain. Too few samples may not allow one a reasonable training and may result in an incorrect query function; too many samples involve a high collection cost. Thus, the number of samples is basically a

tradeoff between cost and reliability of training samples. 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.

In this paper we propose both a novel active learning and a novel semisupervised classification technique based on SVM. Both techniques exploit a specific query function Q which is used in a different way in the two approaches. The query function Q is of fundamental importance in active learning techniques, which often differ only in their respective query functions. The existing methods fall into two broad categories [2]: i) statistical learning approaches, which are designed to minimize expected future errors; and ii) pragmatic approaches, which do not directly consider future performance, but perform a sort of minimization of future errors by selecting the most informative samples for the considered classifier. In this paper we focus on pragmatic approaches for active learning with SVM classifiers.

The proposed active learning technique adopts a novel query function for selecting batches of $h>1$ unlabeled samples at each iteration. This results in speeding up the active learning process with respect to the selection of a single sample at each iteration. To this purpose, the developed query function is based on two basic criteria: i) uncertainty and ii) diversity of samples [4]. The uncertainty criterion, as in standard active learning methods, aims at selecting the least certain samples from the unlabeled sample pool. We evaluate the uncertainty of samples estimating the conditional error by considering the output score of labeled validation samples (the functional distance from the separating hyperplane). Typically, the samples that lie closer to the separating hyperplane of the SVM classifier are considered the most uncertain (and therefore the most informative). However, we have to consider that the position of the hyperplane could be not very precise, especially in the first iterations of the algorithm, where few labeled samples are used for training. Therefore, some randomness is added to the selection of the unlabeled samples. This first criteria is well founded for the selection of the single most informative sample at each iteration. However, selecting a batch of the h most informative samples requires an additional criterion: the diversity of samples. In fact, adding to the training set a batch of new samples that are selected exclusively on the basis of the distance to the separating hyperplane does not necessarily yield a significantly greater information than simply adding a single sample. The diversity criterion aims at selecting a set of unlabeled samples that are as most diverse (distant one to each other) as possible. The combination of the two criteria results in the selection of the potentially most informative h samples at each iteration of the active learning process. Detailed description of the implementation of the two criteria will be reported in the full paper.

Based on the analogy with active learning, we also developed a novel semisupervised classification algorithm that exploits the concept implemented in the aforementioned query function according to a different strategy, but does not require an external supervisor to annotate the unlabeled samples.

3. EXPERIMENTAL RESULTS

We applied both the proposed active learning and semisupervised techniques to binary classification problems associated with very high resolution images acquired by the SPOT and Quickbird multispectral scanners. We compared the proposed active learning technique with other techniques at the state of the art (*least certain* and *confidence based* active learning) and with the *random selection* of new training batches (which serves as a baseline method). Obtained results confirm the effectiveness of the proposed technique that outperformed the other considered methods. Moreover, we observed that active learning allowed us to significantly reduce the computational cost of the classification process, resulting in a very suitable approach for interactive analysis of large scale geographic areas. In addition, the proposed semisupervised technique was compared with other semisupervised algorithms and with the considered active learning methods on the same data sets. The analysis of the obtained results allows one to understand the limits of the semisupervised approach, i.e., in which conditions a semisupervised approach 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 active learning 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.

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] D. Tuia, F. Ratle, F. Pacifici, A. Pozdnoukhov, M. Kanevski, F. Del Frate, D. Solimini, W. J. Emery, "Active learning of very high resolution optical imagery with SVM: entropy vs margin sampling", *Proceedings of IGARSS 2008,* Boston, USA, 2008.

[4] K. Brinker, "Incorporating Diversity in Active Learning with Support Vector Machines", *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, 2003.