MODEL-BASED ACTIVE LEARNING FOR SVM CLASSIFICATION OF REMOTE SENSING IMAGES

Edoardo Pasolli and Farid Melgani

Dept. of Information Engineering and Computer Science, Univ. of Trento,
I-38123 Trento, Italy
E-mail: melgani@disi.unitn.it

1. INTRODUCTION

In general terms, a classification problem consists of associating a sample to a class label chosen from a predefined set of class labels. In the literature, two main approaches to the classification problem have been proposed: the supervised and the unsupervised approaches. The supervised methods are able to give better classification accuracies with respect to the unsupervised ones, but they require the availability of training (labeled) samples for learning the classifier.

For this reason, the performances of supervised strategies depend strongly on the quality and quantity of the labeled data used to train the classifier. Indeed, training samples have to be representative of the statistical distribution of the data. However, the process of collection of training samples is not obvious. Indeed, it is performed by a human expert and thus subject to errors. Moreover, it is costly in terms of both time and money. For this reason, it is necessary to find a strategy to choose few training samples, but fundamental for the correct discrimination between the set of considered classes.

In the last few years, there has been a growing interest in developing strategies for the (semi-)automatic construction of the set of training samples. In the machine learning field, a recent approach focused on this topic is the so-called active learning approach. Starting from a small and suboptimal training set, additional samples, considered important, are selected in some way from a large amount of unlabeled data (learning set). These samples are labeled by the expert and then added to the training set. The entire procedure is iterated until a stopping criterion is satisfied.

In the literature, active learning methods have been applied successfully in different application fields. However, few works have been found for the problem of remote sensing image classification. In [1], the authors present a probabilistic active learning strategy based on support vector machines (SVM) designed for large data applications. Their strategy queries for a set of samples according to a distribution as determined by the current separating hyperplane and an adaptive confidence factor. The confidence factor is estimated from local information using the k-nearest neighbor principle. In [2], a method based on the Fisher information matrix is

used to construct the training set in the application of buried object detection. In [3], the authors propose a probabilistic method based on maximum likelihood classifiers for learning or adapting classifiers when significant changes in the spectral signatures between labeled and unlabeled data are present. In [4], the method proposed in [2] is extended to improve the detection of buried objects. The method fuses a graph-based semisupervised algorithm with an active-learning procedure based on a mutual information measure. In [5], the authors discuss the margin sampling (MS) algorithm [6], a state-of-the-art active learning method based on the SVM classifier. Additionally, two novel methods are proposed and applied to the classification of very high resolution images. The first method is a modification of the MS, which takes into account the distribution of the unlabeled samples in the feature space. In this way, the oversampling on dense regions is avoided and the risk of not sampling important regions is reduced. The second method is independent from the used classifier and is based on a learning committee. A committee of predictors labels the unlabeled samples and the entropy value of the predictions is exploited to decide if selecting or not the considered unlabeled sample.

In this work, an alternative approach of active learning for the SVM classification of remote sensing images is proposed and described in more details in the next Section.

2. PROPOSED METHOD

Let us consider a training set L composed initially of n labeled samples. Each sample has d features (e.g., spectral bands) and is represented by the vector of features $\mathbf{l}_i \in \Re^d = [l_{i,1}, l_{i,2,\ldots}, l_{i,d}]$ *(i = 1, 2, ..., n)* and the corresponding label y_i. y_i assumes one of T discrete values, where T is the number of classes. We consider an additional learning set U, composed of m unlabeled samples $\mathbf{u}_j \in \Re^d = [u_{j,1}, u_{j,2,\ldots}, u_{j,d}]$ *(j = 1, 2, ..., m)*, with $m >> n$.

In order to increase the training set L with a series of samples chosen from the learning set U and labeled manually by the expert, an active learning algorithm has the task of choosing them properly so that to maximize the accuracy of the classification process while minimizing the number of active learning samples to label (i.e., number of interactions with the expert).

The active learning method developed in this work is proposed specifically for classification problems based on SVM. The block diagram of the method is shown in Fig.1. The first step is called significance analysis and consists to detect the most significant samples in the initial training set L. This operation is done by training a multiclass SVM classifier (named SVM1 in the block diagram) on the training set L. We define as significant samples those that the classifier has found as support vectors (SV), while the remaining samples are simply defined as non significant samples. At this point, we construct a new training set L_b, in which the samples of the original training set are relabeled in function of the concept of significance. Therefore, L_b is a binary training set containing significant and non significant samples of L. The successive step of the methodology has the task to estimate the model able to discriminate the significant samples from the non significance samples. For this

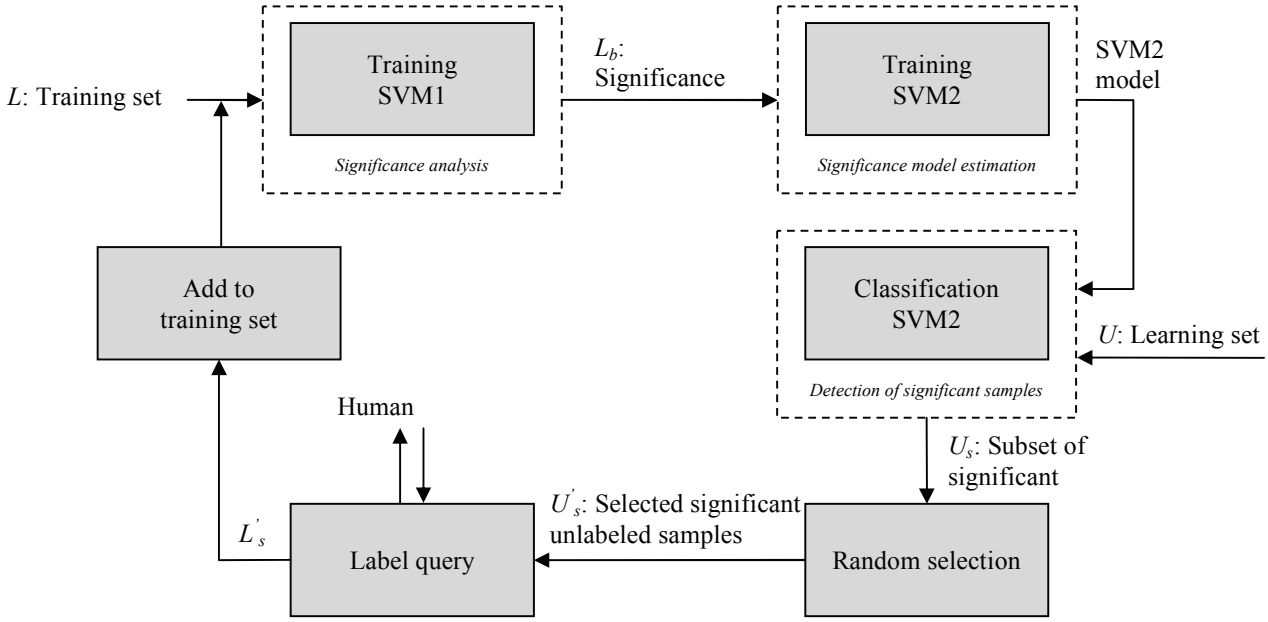

Fig. 1: Block diagram of the proposed method.

purpose, a binary SVM classifier (called SVM2 in the block diagram) is trained on the training set L_b. At this point, we consider the unlabeled samples of the learning set U and estimate their labels using the model defined previously by the second classifier. We define with U_s the samples of the learning set U classified as significant. The last step consists to select randomly N_s samples from the set U_s, where N_s is the number of samples to add to the training set L. Successively, the selected samples U'_s are labeled by the expert and then added to the training set L. The entire process of active learning is iterated until the predefined convergence condition is not satisfied (e.g., the total number of samples to add to the training set is not yet reached).

To validate experimentally our method, we conducted an experimental study based on very high resolution (VHR) remote sensing images. The results were compared with those yielded by a completely random selection strategy, other state-of-the-art active learning algorithms and the "full" training scenario (i.e., a classifier trained on the entire learning set). In general, the obtained experimental results show that interesting performances in terms of accuracy and speed of convergence can be achieved by the proposed method.

3. REFERENCES

[1] P. Mitra, C. A. Murthy, and S. K. Pal, "A probabilistic active support vector learning algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 3, pp. 413–418, Mar. 2004.

[2] Y. Zhang, X. Liao, and L. Carin, "Detection of buried targets via active selection of labeled data: application to sensing subsurface UXO," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 11, pp. 2535–2543, Nov. 2004.

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

[4] Q. Liu, X. Liao, and L. Carin, "Detection of unexploded ordnance via efficient semisupervised and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2218–2232, Jul. 2009.

[5] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 9, pp. 2558–2567, Sep. 2008.

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