Area di ricerca: ICT / ICT
Acronimo: MaleRS
Titolo: Machine learning techniques for the analysis and classification of the last generation of remote sensing data
Data inizio: 18 giugno, 2011
Durata: 36
Finanziamento: € 179.850,00

Abstract 2° anno di progetto

The present research project is about the development of machine learning (ML) techniques for the analysis and classification of the last generation of remote sensing (RS) images. In this project, we focus our attention on the classification of Very High Resolution (VHR) and hyperspectral RS images. Image classification is devoted to automatically recognize the different land-cover types present on the ground and to produce a thematic map of the investigated area. The huge amount of information associated with VHR and hyperspectral RS images makes the classification problem very difficult. One of the main challenges for automatic classification methods is to deal with the high dimensionality of the feature space in consideration of the typical small amount of available labeled training samples. At the state of the art, the most promising techniques for the classification of the last generation of RS data are based on machine learning approaches, like kernel methods and support vector machines (SVMs), which revealed very effective and robust in the solution of many classification problems. For this reason, the general objective of this project is to develop novel machine learning techniques for the analysis and the classification of VHR and hyperspectral images, in order to improve the capability to automatically extract useful information captured from these data and to exploit it in real applications.

In the second year of the project, the research activity has been mainly devoted to 1) the theoretical and experimental analysis of semisupervised (SSL) and active learning (AL) methods for the classification of remote sensing images, and 2) the development of a novel cost-sensitive active learning method that considers the geo-spatial information in the query function to reduce the costs associated with the collection of the training samples. The main idea of SSL is to exploit the structural information of unlabeled samples in the feature space to better model the distribution of the classes and to find a more accurate classification rule than using only labeled samples. AL is an approach to iteratively selecting the
most informative samples to be labeled and used for the training of the supervised classifier by leveraging the feedback from the classifier (the learner).

Regarding the first objective of the research activity, the two learning paradigms have been studied for addressing ill-posed classification problems characterized by a limited amount of training samples and under sample selection bias [1]. This type of problems are very likely to occur in real RS classification problems, especially in the classification of the last generation of RS images, e.g., VHR and hyperspectral images, where the ratio between the available training samples and the number of features is usually small. Commonalities and differences of the two learning approaches have been highlighted in the context of a conceptual framework used to describe the workflow of the two approaches. We pointed out advantages and disadvantages of the two approaches, delineating the boundary conditions on the applicability of the two paradigms with respect to both the amount and the quality of available training samples. Moreover, we investigated the integration of concepts that are in common between the two learning paradigms for improving state-of-the-art techniques and combining AL and SSL in order to jointly leverage the advantages of both approaches. In this framework, we proposed a novel SSL algorithm that improves the Progressive Semi-Supervised Support Vector Machine (PS3VM) by integrating concepts that are usually considered in AL methods. We performed several experiments considering both synthetic and real multispectral and hyperspectral RS data, defining different classification problems starting from different initial training sets.

Regarding the second scientific objective of this second year of the project, we developed a novel cost-sensitive active learning method with lookahead [2, 3]. Standard AL methods typically aim at minimizing the number of labeled samples to be included in the training set to reach a certain level of classification accuracy. Such methods do not take into account the real annotation procedures and implicitly assume that all samples require the same effort to be labeled. We considered the case where the cost associated with the annotation of a given sample depends on the previously labeled samples. In general, this is the case when annotating a queried sample is an action that changes the state of a dynamic system, and the cost is a function of the state of the system. In order to minimize the total annotation cost, the active sample selection problem is addressed in the framework of a Markov Decision Process (MDP), which allows one to plan the next labeling action on the basis of an expected long-term cumulative reward. This framework allowed us to address the problem of optimizing the collection of labeled samples by field surveys for the classification of remote sensing data. The proposed method has been applied to the ground sample collection for tree species classification using airborne hyperspectral images. Experiments carried out in the context of a real case study on forest inventory show the effectiveness of the proposed method.

References

