Active Learning classification for spatially and auto-correlated data: Application to remote sensing data

Phd Supervisors : Dr. Dino IENCO and Pr. Maguelonne TEISSEIRE
{ienco, teisseire}@teledetection.fr

Context

The study of the spatial dynamics of regional or national scales leads to cross multi-source data of various types maps, surveys, sensor data or satellite images.
These data carry information of spatial, temporal and thematic character. The study of satellite images allows automated explorative analysis of the territories. Once the images are acquired, a process of segmentation is performed to extract the objects of interest. The segmentation step is a critical phase that requires a certain expertise, both on the tool used and the nature of the land concerned. This phase is a key to the success of the analysis of the land remote sensing process. Once the segmentation carried out, a set of objects is obtained, whose number depends on the granularity associated with the study to be conducted. The ultimate goal is to obtain a classification of these objects to produce as close as possible to the reality of land use mapping.
The number of objects produced during segmentation can be huge (between 5000 and 50000 objects), for this reason it is not always possible to ask the experts to validate all these items manually. This problem of Object Oriented Classification (OOC) involves the automatic classification of these segment obtained by the satellite image in order to produce a detailed map of the study area. In order to automatically perform this operation we need to train a predictive model with a labeled training set [5]. Most of the time, these methods assume that a big set of example, already labeled, exist and it is easily ready to be employed.
The construction of this training set has a cost, as the expert must analyze each object one by one and assign a class of land. Once this step is completed, the classification algorithm can then be driven [6].
Today, the software reference tool in the field of object oriented classification remains eCognition which offers several automatic methods for classifying objects obtained from a satellite image. Classification methods proposed were generally developed for applications from the biomedical domain imaging (health). These classification approaches make the assumption that objects are completely independent of each others (iid = identically distributed Independently) while in the satellite imagery the spatial neighborhood can be informative to help classify a specific object. This phenomenon is called spatial autocorrelation [4], i.e. the objects are influenced by neighboring objects.

Objectives of the thesis

Starting from all the previous aspects, during this thesis we would to focus the PhD efforts to improve current techniques for the OOC of satellite images considering active learning techniques to reduce the cost of the labeling process and involving spatial information that really characterized our domain application.

The three main points of this thesis are:

1) Active Learning:

We want to use and study active learning techniques [1] to select the best training set by reducing its cost (that is to say, the effort required to the experts). Active learning is a paradigm that pose constraints on the amount of available data in training phase. Instead of asking the expert to classify a random collection of objects, the system will select automatically and intelligently, the objects to be labelled to minimize effort and maximize the annotation result of the classification. In particular, the system, employing some heuristics, selects example to supply to the experts who will label them to build an useful training data to build the predictive model [2]. Normally in an active learning process, the user chooses a percentage (the concept of "Budget") which indicates how many objects (maximum) the system can ask the expert to label.
The advantages of such an approach is to build a classifier by minimizing the number of objects to be labelled.

2) Neighborhood space:

We want to take into account the fact that the objects classified are spatially correlated. To manage this phenomenon, it is possible to act at two levels. The first is to integrate this aspect in the active learning process. In particular, we can study and develop dedicated active learning methods that do not follow the iid hypothesis [3].
The request to provide objects to be labelled by the expert can then integrate the spatial dimension avoiding asking nearby objects or otherwise using this spatial proximity in order to get more information.
The second possible point of view is to develop (or adapt) directly classification algorithms. At this level, a number of efforts have been made to develop methods for i.i.d. but few studies have addressed the issue of auto-correlated data [4] and even less for extracting knowledge from self-spatially correlated data.
In order to deal with this aspect we can employ concepts coming from geostatistics (range correlation variogram, etc ...) or one-time spatial processes (attraction / repulsion of objects) to try to model the spatial autocorrelation between land cover objects.

3) Time evolution:

The last point of the thesis will focus on the temporal dimension to improve the process of OOC. All methods of Object Oriented Classification does not consider time series of images while become important to rely on temporal trends in order to identify patterns of object behaviours [7] at pixel level since most the current work is based on a pixel approach. [10]

Application data

Ideal applications are two areas on which TETIS team - Irstea has already invested:

1) Monitoring the phenology and mapping of natural habitats
The Natura 2000 site of the Lower Plaine de l'Aude is currently a study site from our team in TETIS. Its uniqueness lies in the close integration of natural environments, depending on

the conditions imposed first by salinity of water and topography but also by the quality of soil, climatic characteristics, and human activities. He is a representative environment for testing new methodologies proposed in this thesis. The multi-date images are available and a ground truth carried out by CEN is available.

2) Monitoring of urban evolution
The continued growth of artificial surfaces spaces at the expense of natural, agricultural and forestry products, and the recent strengthening of the regulatory environment to contain space consumption (Law for the Modernization of Agriculture and Fisheries 2010, Grenelle 2 2010 SRU 2005) require the development of methods for locating and quantifying artificialized spaces, applicable nationwide. The conventional sensing methods can satisfy this need. However, their implementation across the entire French territory requires significant human resources. The proposed methodology represents a potential pathway of development by integrating new approaches to improve the relevance of results and automated processing.
So it would assess the complementarity of the proposed methodologies with traditional remote sensing technics.

Bibliographie

[1] D. A. Cohn, Z. Ghahramani and M. I. Jordan. Active Learning with Statistical Models, Journal of Artificial Intelligence Research, JAIR, 4: 129-145 (1996)
[2] H. T. Nguyen and A. W. Smeulders, Active learning using pre-clustering. ICML 2004
[3] M. Dundar, B. Krishnapuram, J. Bi and R. B. Rao. Learning Classifiers when the training data is not IID, IJCAI (2007)
[4] D. Stojanova, M. Ceci, A. Appice, D. Malerba and S. Dzeroski. Dealing with spatial autocorrelation when learning predictive clustering trees. Ecological Informatics, 13:22-39 (2013)
[5] T. M. Mitchell, Machine Learning, McGraw-Hill, 1997
[6] A. Blum and T. M. Mitchell. Combining Labeled and Unlabeled Data with co-Training. COLT (1998)
[7] E. Vintrou, D. Ienco, A. Begue and M. Teisseire. Data Mining, a promising tool for large area cropland mapping. IEEE Selected Topics in Applied Earth Observations and Remote Sensing, Accepted for publication, 2013
[8] Andrienko G., Malerba D., May M., Teisseire M. (2006). Special issue . Mining Spatio-Temporal Data. of JIIS Journal of Intelligent Information Systems, Kluwer Academic Publishers, Volume 27, Number 2, September 2006.
[9] Hugo Alatrista Salas, Sandra Bringay, Frédéric Flouvat, Nazha Semaloui and Maguelonne Teisseire "The Pattern Next Door: Towards Spatio-Sequential Pattern Discovery" In Proceedings of the 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2012), Lecture Notes in Artificial Intelligence (LNAI), Kuala Lumpur, Malaysia, May-June 2012, pp. 157-168.
[10] F. Petitjean, C. Kurtz, N. Passat, P. Gançarski Spatio-temporal reasoning for the classification of satelitte image time series. Pattern Recognition Letters, pp. 1805--1815, Vol. 33, Num. 13, doi:10.1016/j.patrec.2012.06.009, October 201