Proposition for PhD 2021 – 2024 in Paris

Title	Multiparadigm interactive collaborative learning for heterogeneous remote sensing time series analysis
Laboratories	**Mathématiques et Informatique Appliquées - UMR 518 AgroParisTech, INRAE, Université Paris-Saclay** Laboratoire ICube, UMR CNRS 7357, Université de Strasbourg,
Academic supervisors	**Cornuéjols Antoine** – 06 31 65 25 66 - antoine.cornuejols@agroparistech.fr Gançarski Pierre – 06 87 44 58 50 - gancarski@unistra.fr
Funding	ANR Herelles (Doctoral contract)
Profile of applicant	Master's Degree in Computer Science. The candidate must have good skills in data analysis and more particularly in supervised or unsupervised classification of time series. Skills in remote sensing image analysis is required. Good knowledge of English (French is not mandatory)

1 SUMMARY

Analysing heterogeneous remote sensing time series using supervised methods requires that the classes sought are perfectly known and defined and that the expert is able to provide a sufficient learning data set both in number and quality. Faced with the difficulty of obtaining sufficient examples within the context of the analysis of time series of remote sensing images, we propose to develop an innovative method of interactive multi-paradigm collaborative learning. The aim is to enable the expert to add "on the fly" information (labels, constraints, etc.) used to guide the learning process in order to produce clusters and models closer to the expert's "intuitions", i.e. potential thematic classes. To do this, the expert will be actively assisted by the system, which will for example offer advice or proposals for new constraints or labelling of objects. We will validate our work in several fields of application chosen in agreement with partners of the HERELLES project.

2 THESIS SUBJECT DESCRIPTION

2.1 CONTEXT AND OBJECTIVE

With the launch and entry into production of the European satellites in the Sentinel or Franco-Israeli constellation Venųs, satellite data are now arriving in massive, almost continuous flows. This massive influx of temporal data should lead to major advances in various Earth and environmental science disciplines for the study and modelling of complex phenomena (agricultural or urban dynamics, deforestation, anthropogenic actions on biodiversity, etc.). However, faced with this overabundance of temporal data, arriving almost continuously, the labelling phase of supervised learning can no longer be carried out by experts, as it is too tedious and time-consuming. Moreover, the supervised learning methods classically used in Earth observation assume that the learning data sufficiently and completely describe the classes to which they are attached. In other words, these methods require that the desired classes are well known and defined and that the expert is able to provide a sufficient set of learning data both in number and quality. In the case of temporal analysis in remote sensing, this assumption is no longer realistic. Indeed, the technological revolution of high-frequency image acquisition is still too recent for thematic knowledge to have adapted. Thus, there are currently no typologies (or nomenclatures) of changes that can really be used for this type of supervised analysis and therefore no associated quality learning data.

To compensate for this lack of formalization and examples, the expert must be able to rely on other types of information such as partially labelled data, formalized knowledge, constraints on data or results. At the same time, there are also numerous methods capable of analyzing this data. Combining these data and methods seems indispensable. Thus, approaches such as boosting [13], clustering [14] or collaborative clustering [2] take advantage of the complementarity between different methods, each with its own biases and its own analysis strategy but capable of processing its own data in a privileged way.

However, with the increase in the volume of data and the number of potential evolution classes, the highlighting and formalization of information that is really relevant for classification methods in the context of temporal analysis appears to be more difficult than expected and potentially time-consuming. The objective of this project, in strong link with the ANR HIATUS and HERELLES projects, is to define and validate in the context of high acquisition frequency remote sensing, *an innovative method of interactive collaborative learning*. The aim is to enable the expert to add "on the fly" information (labels, classes, constraints, etc.) that can be used to guide the learning process in order to produce clusters and models that are closer to the expert's "intuition", i.e. potential thematic classes [1,3,4,11,12]. To do this, the expert will be actively assisted by the system, which will offer advice or proposals for new constraints or object labelling, for example.

2.2 SCIENTIFIC ASPECTS

Interactive collaborative learning. Selecting the new information (object to be labelled, new constraint to be applied...) that is relevant, i.e. that has a positive impact on the current result, is often very difficult for the expert. Indeed, the expert almost exclusively uses a visualization of the scene. Experiences have shown that on the one hand, the expert focuses on relatively large areas of the image and that on the other hand, he has no way of knowing whether the information he proposes is consistent with each other and relevant a priori. In fact, selecting new information is an important scientific hurdle, all the more so as it is essential to optimise the use of this new information coming from the expert. Indeed, if the expert does not see a rapid improvement in the solution as a result of his help, he will quickly lose confidence in the system. But, paradoxically, potential disruptions to the current solution must be limited so as not to disorientate the expert.

In our collaborative framework, information can be considered either at the global level of the meta-classifier or locally by each classifier. At the global level, the differences between the results due to the heterogeneity of the methods can be exploited to find potentially interesting information. For example, to find new constraints, we will draw on the so-called global methods in order to find a consensus and thus identify the "sensitive" elements with, for example, low agreement [5,6]. At the local level, one approach envisaged is to use the data complexity measures proposed by the supervised methods such as, for example, a complexity measure based on trees of minimum weight to identify the points at the boundaries between clusters and use them to define constraints. This complexity measure is related to the internal criterion of informativeness which often takes the form in supervised learning, of max-margin sampling, where samples are selected according to the maximum uncertainty across distances from the classification boundary [7,8]. Another more disruptive approach would be to rely on a history of expert feedback [10] to try to identify the most informative information and assess its impact on the model and thus limit its disruption.

2.3 APPLICATIVE ASPECTS

For the consolidation of the proposals and the thematic validation, the PhD student will be able to rely on the work undertaken and the interactions set up with SERTIT in the context of the A2CNES R&T contract and with IGN in the one of the ANR HIATUS project. The collaborations thus initiated will continue throughout the thesis in order to allow qualified feedback on the methods developed. Different fields of application (chosen in agreement with the CNES) are envisaged. They will mainly concern situations in which the types of evolution are both numerous and not very formalized. By way of example (not exhaustive), we may be interested in the following:

- The different forms of evolution of urban elements (buildings, green spaces, infrastructures, etc.).
- Monitoring the (re)vegetation in the vicinity of new infrastructures: this will involve highlighting the classes of revitalization/resettlement of vegetation in the vicinity of newly created infrastructures, then monitoring the evolution of this vegetation over several years.
- In the aftermath of disasters, i.e., the possible changes in natural re-greening or the reconstruction of buildings and infrastructure in order to deduce "resilience" classes.
- The analysis of past phenomena such as the different processes of wind turbine installation.

- Discovering the different methods of cutting in the forest massifs: The detection of clearcuts has already been the subject of previous studies. The case of selective cutting, which is much more complex, can be addressed.

The use of data combining images or heterogeneous series with different modalities in terms of acquisition frequencies as well as spatial, spectral and/or radiometric resolutions, authorised by the collaborative approach should allow a better understanding of the phenomena underlying these images in their entirety.
The work of this thesis will be integrated to the FoDoMuST-MultiCube [9] platform dedicated to the multitemporal analysis of remote sensing data. It should be noted that within the context of the A^{2}CNES R&T contract, this platform has been made compatible with the OTB.

2.4 SUPERVISION AND COLLABORATION

The person recruited will be supervised by Antoine Cornuéjols (AgroParisTech - 50%), a specialist in collaborative supervised learning and Pierre Gançarski (ICube - 50%), a specialist in collaborative clustering. Thus, thanks to the potential interaction with CNES (via Emmanuelle Sarrazin) all aspects of the project will be covered: supervised and unsupervised collaborative learning and remote sensing.

The work will also benefit from a strong collaboration with Hussein El Amouri, PhD student (2nd year), and Antoine Saget (PhD starting in October 2021) working both on the extraction (from) and use of constraints (in) collaborative clustering of time series (ANR HIATUS and ARTIC projects).

3 RÉFÉRENCES BIBLIOGRAPHIQUES

[1] T. Lampert, T-B-H. Dao, B. Lafabregue, N. Serrette, G. Forestier, B. Crémilleux, C. Vrain, P. Gançarski, Constrained distance based clustering for time-series: A comparative and experimental study. Data Mining Knowledge Discovery, 32:1663–1707 (2018)

[2] P. Gançarski, C. Wemmert, Collaborative Multi-step Mono-level Multi-strategy Classification, Multimedia Tools and Applications, Springer 35(1) pages 1-27, 2007

[3] B. Sugato, I Davidson, K. Wagstaff "Constrained Clustering: Advances in Algorithms, Theory, and applications", CRC Press

[4] D. Derya M. K. Tural, "A Survey of Constrained Clustering" In book: Unsupervised Learning Algorithms, pp.207-235.

[5] S. Vega-Pons, J. Ruiz-Shulcloper. A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence, 25:337–372 (2011)

[6] O. Sagi, L. Rokach. Ensemble learning: A survey. Data Mining and Knowledge Discovery, 8 (2018)

[7] B. Du, Z. Wang, L. Zhang, L. Zhang, W. Liu, J. Shen, D. Tao. Exploring representativeness and informativeness for active learning. IEEE Transactions on Cybernetics, 47:14–26 (2017)

[8] M. Wang, X.-S. Hua. Active learning in multimedia annotation and retrieval: A survey. ACM Transactions on Intelligent Systems and Technology, 2:1–21 (2011)

[9] http://icube-sdc.unistra.fr/en/index.php/FODOMUST

[10] P. Daee, T. Peltola, M. Soare, S. Kaski. Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. Machine Learning, 106:1599–1620 (2017)

[11] I. Kopanas, N. M. Avouris and S. Daskalaki, The role of domain knowledge in a large scale Data Mining project. Methods and Applications of Artificial Intelligence: LNAI pages 288-299

[12] S. S. Anand, D. A. Bell, et J. G. Hughes. The role of domain knowledge in data mining. In CIKM '95: Proceedings of the fourth international conference on Information and knowledge management, pages 37–43, New York, NY, USA, 1995. ACM. ISBN 0-89791-812-6.

[13] R.E. Schapire et al. Boosting: Foundations and Algorithms. MIT Press (2012).

[14] J. Piantoni et al. Impact of base partitions on multi-objective and traditional ensemble clustering algorithms. In ICONIP, pp. 696–704 (2015).