Proposition for PhD 2020

Title	Interactive collaborative constrained clustering for remote sensing time series analysis
CNES supervisor	Victor Poughon – Service DCT/SI/AP CNES
Laboratory	Laboratoire ICube, UMR CNRS 7357, icube.unistra.fr
Academic supervisor	Professor **Gançarski Pierre** - 03 68 85 45 76 - gancarski@unistra.fr
Project	ANR Hiatus (Garanti)
Profile of applicant	Master 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 ABSTRACT

Analyzing time series of remote sensing images using supervised methods requires that the classes sought be perfectly known and defined and that the expert be able to provide a learning data set that is sufficient in number and quality. Facing the difficulty of obtaining sufficient examples to efficient remote sensing time series analysis, we propose to develop an interactive method of collaborative clustering under constraints. The idea is to allow the expert to add "on the fly" constraints to guide the clustering process in order to produce clusters closer to the expert's "intuitions" i.e. potential thematic classes. To do so, the expert will be helped by advice or proposals for new constraints issued by the method itself.

2 PHD TOPIC

2.1 CONTEXT

Nowadays, remote sensing images arrive massively and in almost continuous flow from the Sentinel constellation. 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.). Analyzing such time series using classical supervised methods requires that the thematic classes are perfectly known and defined. Unfortunately, in remote sensing domain, this assumption is not 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.

Faced with the difficulty of obtaining enough examples for the analysis of time series of remote sensing images, new clustering methods used constraints to guide the clustering process [1,3,4,5]. These unsupervised methods assume that this lack of knowledge can nevertheless be partially circumvented by using operable constraints (comparison, labelling or structural constraints). Such as constraints, which seem easier to define, can be used to guide the clustering process in order to produce clusters that are closer to the "intuitions" of the expert, i.e. potential thematic classes. In our team, we have developed SAMARAH an innovative method of collaborative interactive clustering under constraints [2] which allows the expert to add "on the fly" constraints.

Unfortunately, select new relevant constraints (object to be labelled, new constraint to apply...) that have positive impact on the current result, is often very difficult for the expert. Indeed, to define new constraints, the expert uses almost exclusively a visualization of the scene. Experiences have shown that, on the one hand, the expert focused on relatively large regions of the image and, on the other hand, he had no way of knowing whether the constraints he proposed were consistent with each other and relevant a priori. In fact, selecting new information is an important scientific lock, especially since it is essential to optimize the use of this new information from the expert. Indeed, if he does not see a rapid improvement of the solution following his help, he will quickly lose confidence in the system. But, paradoxically, the potential disruptions of the current solution should be limited in order not to disorient the expert. To this end, the

expert must be assisted by advice or proposals for new constraints issued by the method in an active way [6,7].
The objective of this PhD is to study and implement mechanisms to propose potentially relevant constraints. This can be done, for example, by using differences in results due to the heterogeneity of methods in SAMARAH or by using a complexity measure, for example, based on trees of minimal weight to identify points at the boundaries between clusters and use them to define constraints.

2.2 CONSOLIDATION OF PROPOSALS AND THEMATIC VALIDATION

For the consolidation of proposals and thematic validation, the internship student will be able to rely on the work undertaken and the interactions set up with SERTIT. Different fields of application are envisaged such as for example (in a non-exhaustive way):

1. Detection and monitoring of cuts in the Vosges mountains: The detection of clear cuts has already been the subject of previous studies. The case of selective cutting, which is much more complex, could be studied.
2. Monitoring of (re)vegetation around new infrastructure: this will involve identifying vegetation revitalization/reinstallation classes around newly created infrastructure and then monitoring the evolution of this multi-annual vegetation.

The proposed mechanism(s) will be integrated into the FODOMUST-MULTICUBE platform [8] dedicated to the multi-temporal analysis of remote sensing data.

2.3 COLLABORATIONS

The contributions of the PhD student will be part of the global work of the SDC researchers who aim to propose and implement new generic methods and tools to analyze massive time series. Collaborations will take place with the SERTIT platform of the ICube laboratory for the provision of pre-processed time series via the A2S platform if necessary, and for the provision of thematic expertise. Thomas Lampert and Robin Faivre will participate in the supervision of the thesis.

The candidate (with a Master in Computer Science) 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 welcome.

Location: ICube, Strasbourg, France

To apply, send an email to Professor Pierre Gançarski (gancarsi@unistra.fr) *and* Dr Thomas Lampert (lampert@unistra.fr) be sure to include your C.V. and a cover letter to let us know why you think you would be a good fit.

The candidate has also to apply on the CNES website (Button on the left) before March 31:
https://recrutement.cnes.fr/en/annonce/898696-70-interactive-collaborative-constrained-clustering-for-remote-sensing-67400-illkirch-graffenstaden

3 REFERENCES

[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.