Proposition for internship for 2020

Subject	Integration of a constraint extraction mechanism into a collaborative clustering process
Laboratory	Laboratoire ICube, UMR CNRS 7357, icube.unistra.fr
Supervisors	**Pierre Gançarski** gancarski@unistra.fr **Thomas Lampert** lampert@unistra.fr
Candidate profile	Master of Computer Science. Skills in data mining and satellite imagery.

1 OBJECTIVE

Analysing satellite image time-series using supervised methods requires that thematic classes are perfectly known and defined, and that the expert is able to provide a sufficient set of training data in terms of both number and quality. Faced with the difficulty of obtaining enough examples for the such an analysis, new clustering methods use constraints to guide the clustering process [1,3,4,5]. In particular, in our team, we have developed SAMARAH an innovative method of collaborative interactive clustering under constraints [2]. This method allows the expert to add constraints "on the fly" to guide the process in order to produce clusters closer to the expert's "intuition", i.e. potential thematic classes. Thus, the SAMARAH collaborative method developed by ICube allows constraints to be considered incrementally.

Nevertheless, selecting which piece of additional information (object to be labelled, new constraint to apply, etc) is most relevant, i.e. that has a positive impact on the current result, is often very difficult for the expert. Indeed, to define new constraints, the expert almost exclusively uses a visualisation of the scene. Experiments show that, on the one hand, the expert focus on relatively large regions of the image and, on the other hand, they have no way of knowing whether the constraints that are proposed are consistent with each other and relevant a priori. In fact, selecting new information is an important scientific problem, especially since it is essential to optimise the manner in which to obtain this new information from an expert. If they do not see a rapid improvement of the solution following their help, they will quickly lose confidence in the system. Paradoxically, the potential disruptions to the current solution (by the new information) should be limited in order not to disorient the expert. To this end, the expert must be assisted with advice or propositions for new constraints by the method in an active way [6,7].

The objective of this internship is to study and implement mechanisms to propose potentially relevant constraints. This can be done, for example, using two approaches [1]: dependent on, and independent from the clustering algorithm. Ideas in the algorithm dependent direction are, to use the difference between results due to the heterogeneity of methods in SAMARAH, and/or by developing new measures based on the inconsistency [8] and informativeness [9] measures. Directions in the algorithm independent direction are to use 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, or by developing new measures similar to coherence [9] for time-series.

2 CONTEXT

For the consolidation of proposals and thematic validation, the intern will be able to rely on the work undertaken between ICube and SERTIT. Different fields of application are envisaged such as (non-exhaustively):

1. Detection and monitoring of tree 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 revitalisation/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 [10] dedicated to the multi-temporal analysis of remote sensing data.

The candidate, 2nd year student of a Master's of Computer Science degree, must have good skills in data analysis and more particularly in supervised or unsupervised classification of time series. Skills in remote sensing image analysis are welcome.

Place: ICube – Pôle API, Illkirch.

Salary: 550 € per month.

Send your CV, transcript of grades, and motivation to Pierre Gançarski and Thomas Lampert.

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):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] 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)

[7] 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)

[8] K. Wagstaff, S. Basu, I. Davidson, When is constrained clustering beneficial, and why? In: Proceedings of the National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, (2006).

[9] I. Davidson, S. Ravi, Identifying and generating easy sets of constraints for clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 336–341, (2006).

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