PhD position in Artificial Intelligence

In the context of a **national project HERELLES**, supported by ANR (Agence Nationale de la Recherche), we offer a PhD position at LIFO, University of Orléans in collaboration with GREYC/LS2N, University of Caen Normandy. The PhD will start in October 2021 for 3 years.

Related to the field of **Artificial Intelligence,** the PhD research will include the following topics: Machine Learning, Data Mining, Constraint Programming and Applied Mathematics. The Thesis title is:

Constrained clustering: incremental and active integration

PhD position specifications

- The PhD position is fully funded by the ANR Project HERELLES.
- The PhD position will be conducted at **LIFO, University of Orléans** in collaboration with GREYC/LS2N, University of Caen Normandy.
- The PhD position will start in **October 2021**.
- **The deadline to apply is 23rd of May 2021**
- The audition will be in early June 2021.

Required technical skills

- Ability to communicate effectively in French or in English, both orally and in writing.
- Strong skills in programming languages such as C++, Java and Python.
- Experience in machine learning, data mining, constraint programming. Skills in applied mathematics are highly appreciated.

Required documents to apply

The complete application consists of the documents below, which should be sent as a single PDF file to:

Thi-Bich-Hanh Dao (thi-bich-hanh.dao@univ-orleans.fr)
LIFO, University of Orleans

Samir Loudni (samir.loudni@imt-atlantique.fr)
IMT Atlantique - CNRS - LS2N

- Detailed CV
- One-page cover letter (clearly indicating available starting date as well as relevant qualifications, experience and motivation)
- University certificates and transcripts (both B.Sc and M.Sc degrees marks)
- Contact details of up to three referees
- Possibly an English language certificate and a list of publications
- Attention: all documents should be in English or in French.

PhD Subject Description

Clustering is an important task in Data Mining, which aims at partitioning data instances into groups to find the underlying structure of the data. Clustering has been extended to constrained clustering, which allows to integrate prior expert knowledge in the form of constraints, in order to make the clustering task more accurate [1]. Most constrained clustering methods request the specification of all the constraints before the subsequent running of the methods. It is however crucial that the expert could interact with the clustering process, that he could inject new information or knowledge in the form of constraints on a clustering result. Constraints can be pairwise must-link or cannot-link constraints, which state that two instances must be or cannot be in the same cluster, or can be constraints on the clusters, stating bounds on their size or their diameter, or can be operations on clusters, such as split a cluster or merge two clusters, etc. The constrained clustering process therefore becomes incremental and interactive. However, these two properties are not well considered in existing constrained clustering approaches. This thesis aims to investigate these two research directions.

The thesis will be organized into two complementary parts. The objective of the first part is to propose new clustering approaches enabling to integrate incrementally new constraints identified as important by the expert or by measures, such as [6]. Declarative approaches based on Constraint Programming (CP) or Integer Linear Programming (ILP) will be considered due to their expressiveness and their rich constraint language. Meanwhile, in order to avoid confusing the expert, the new partition solution should not be too different from the previous one. This could be guaranteed based on a measure of clustering similarity, which can be either statistical [8] or more explanatory [5].

In the second part, we will consider a more user-centered and interactive clustering approach. This new paradigm stresses that users should be presented quickly with new generated constraints likely to be interesting to them (i.e., which may improve clustering quality in later iterations), by giving feedback. These feedback could be of the form validate / invalidate the constraints. In the context of mono-clustering, these constraints can be generated based on the information on an existing partition to identify informative points (e.g. frontier points). We will also consider the case where a set of clusterings is available, like in collaborative and multiparadigme clustering [2,7]. In such settings, one can use information from different clusterings to identify for instance uncertainty pairs or to elicit best objective functions according to some criteria to be defined. Here, a pair is more uncertain if more clusterings disagree on whether it should be in the same cluster or not. Another approach is to exploit the history of the feedback to determine most informative points [3]. Meanwhile, to prevent contradiction during the collection of the user feedback, consistencies on the learned constraints must be ensured.

The proposed method will be generic and will not depend on the potential areas of application. As part of the HERELLES project, in order to validate the operability of the method, we will focus on understanding complex phenomena in our environment (soil artificialization, urbanization, construction of infrastructure, etc.) mainly via heterogeneous temporal data.

References:

[1] S. Basu, I. Davidson, K. Wagstaff. Constrained Clustering: Advances in Algorithms, Theory, and Applications. CRC Press (2009)

[2] A. Cornuéjols, C. Wemmert, P. Gançarski, Y. Bennani. Collaborative clustering: Why, when, what and how. Information Fusion, 39:81–95 (2018)

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

[4] T.-B.-H. Dao, K.-C. Duong, C. Vrain. Constrained clustering by constraint programming. Artificial Intelligence 244: 70-94 (2017)

[5] T.-B.-H. Dao, C. Kuo, S.S. Ravi, C. Vrain, I. Davidson. Descriptive clustering: ILP and CP formulations with applications. IJCAI, pp. 1263–1269 (2018)

[6] I. Davidson, K. Wagstaff, S. Basu. Measuring Constraint-Set Utility for Partitional Clustering Algorithms. PKDD 2006: 115-126

[7] G. Forestier, P. Gançarski, C. Wemmert. Collaborative clustering with background knowledge. Data & Knowledge Engineering, 69:211–228 (2010)

[8] C. Kuo, S. S. Ravi, T.-B.-H. Dao, C. Vrain, I. Davidson. A Framework for Minimal Clustering Modification via Constraint Programming. AAAI 2017: 1389-1395

[9] N.-V.-D. Nghiem, C. Vrain, T.-B.-H. Dao, I.Davidson. Constrained Clustering via Post-Processing, in 23rd International Conference on Discovery Science. 2020.

[10] A. Ouali, S. Loudni, Y. Lebbah, P. Boizumault, A. Zimmermann, L. Loukil. Efficiently Finding Conceptual Clustering Models with Integer Linear Programming. IJCAI 2016: 647-654