Stage recherche de 5 à 6 mois

Clustering Ensemble sous Contraintes

LIFO- Université d'Orléans

Contact : Christel Vrain

christel.vrain@univ-orleans.fr

Encadrants : Christel Vrain, Marcilio Pereira de Souto, Thi Bich Hanh Dao

Ce stage recherche est lié à un projet national InvolvD, financé par l'ANR (Agence National de la Recherche) débutant en Février 2021. Ce projet comporte aussi une bourse pour une thèse dont l'appel à candidature sera publié au Printemps 2021.

La classification non supervisée (clustering) a pour but de trouver des structures sous-jacentes présentes dans les données, comme par exemple une partition (clustering) des données en groupes. Les observations appartenant à un même groupe doivent alors partager des propriétés pertinentes par rapport à l'application visée. Intégrer des connaissances du domaine peuvent permettre de guider le processus vers un clustering, plus proche des besoins de l'expert. Elles peuvent porter sur des paires de points exprimant que deux points doivent, resp. ne doivent pas être dans le même cluster, ou des contraintes sur les clusters (par exemple leur taille ou leur diamètre). Cela a conduit à un nouveau courant de recherche appelé Clustering sous Contraintes. De nombreuses méthodes ont déjà été développées pour intégrer des contraintes dans un processus de clustering. Certaines sont dédiées à un type de contraintes, d'autres sont plus génériques, souvent fondées sur des cadres déclaratifs comme la Programmation Linéaire en Nombres Entiers, la Programmation par Contraintes ou SAT.

Au lieu de produire un unique clustering sur lequel l'utilisateur peut donner un avis (feedback), on peut lui présenter plusieurs partitions et le laisser choisir des clusters qui lui semblent pertinents ou proposer la fusion de clusters qui partagent des propriétés similaires. Dans ce stage nous nous intéressons à l'intégration des retours de l'expert en présence de plusieurs partitions construites. A ces fins, nous devons développer deux aspects :

1) Interprétabilité: nous sommes intéressés par des applications en chemo-informatique où les données sont représentées par des descripteurs discrets. Pour faciliter la tâche de l'expert, nous devons développer des approches qui mettent en évidence les différences/similarités entre couples de clusters et ainsi proposent des interprétations des clusters, dont le niveau dépend de la connaissance structurelle ou sémantique disponible.

2) Fusionner différents clusters sous contraintes données par l'expert. L'idée est qu'il existe plusieurs partitions satisfaisant partiellement l'exert et qu'elles doivent être fusionnées dans une partition consensus satisfaisant toutes les contraintes. Nous considèrerons des méthodes purement déclaratives garantissant de trouver une partition consensus satisfaisant toutes les contraintes.

Ce stage de recherche a pour but de

- Produire un état de l'art sur les méthodes de clustering ensemble sous contraintes utilisateurs
- Proposer des explications, étant donné un ensemble de partitions
- Proposer et tester un premier prototype de clustering ensemble sous contraintes.

Compétences demandées :

- Compétences en machine learning/data mining. Bonnes capacités en programmation. Des connaissances en Programmation par Contraintes seraient appréciées.

- Une maîtrise du Français et/ou de l'Anglais

Les candidats sont encouragés à nous contacter aussi vite que possible. Le début du stage est prévu en Février 2021. Pour postuler les documents suivants doivent être envoyés en un seuf fichier pdf à Christel Vrain (christel.vrain@univ-orleans.fr)

- CV
- Lettre de motivation d'une page (indiquant clairement la date de début, les compétences ainsi que la motivation pour ce stage)
- Relevés de notes de l'Université (Licence et Master)
- Contacts de 3 personnes référentes
- Attention: tous les documents doivent être en anglais et en français.

========== English version ==========

5 or 6-month research internship

Ensemble Constrained Clustering.

LIFO - University of Orléans - France

Contact : Christel Vrain

christel.vrain@univ-orleans.fr

Supervisors : Christel Vrain, Marcilio Pereira de Souto, Thi Bich Hanh Dao

This master internship is part of a national project InvolvD, supported by ANR (Agence Nationale de la Recherche), starting in February 2021. This project also includes a grant for a PhD ; the call for application will be published in Spring 2021.

Clustering is a type of unsupervised learning whose goal is to find the underlying structure present in the data as, for examples, a partition/clustering composed of groups/clusters. Observations belonging to each cluster should share some relevant property (similarity) regarding the data domain. Integrating knowledge can help guiding the process toward a clustering, closer to the expert needs. It can be pairwise constraints, such as must-link or cannot-link constraints, expressing that two points should be, resp. cannot be, in the same cluster, or constraints on the clusters (for instance their size, diameter, ...) This has led to a new research area called Constrained Clustering and many methods have already been developed for integrating constraints in a clustering algorithm. Some of them are dedicated to one kind of constraints, others are generic, usually based on declarative frameworks such as Integer Linear Programming, Constraint Programming, SAT.

Instead of a single clustering on which the user can give a feedback, one can present her several partitions and let her pick only one or merge those that show desired characteristics in parts of them. In this internship, we are interested in **integrating feedback given by an expert on several partitions**. To do this, we will need to develop two aspects.

1) Interpretability: we are interested in applications (chemo-informatics) where data is represented by discrete descriptors. To ease the expert in giving relevant feedback we will develop approaches that highlight the differences/similarities between pairs of clusters and thus provide interpretations, the level of explanations depending on the structural or semantic information available.

2) Merging different clusterings under constraints given by the expert. There might not be a single preferred clustering but several ones that should be merged into a consensus partition while satisfying certain constraints. We will consider purely declarative methods that guarantee to find a consensus partition satisfying all the constraints.

This internship aims specifically at:

- Producing a review of the state-of-the-art on ensemble methods and on constrained ensemble clustering
- Proposing explanations on the multiple clusterings
- Proposing and testing novel (or improved) constrained ensemble methods

Required skills:

- Experience in machine learning, data mining, computer programming or applied mathematics is highly appreciated.

- French and/or English are the working languages.

Candidates are encouraged to contact us as soon as possible. Start is expected on February 2021. The complete application consists of the documents below, which should be sent as a single PDF file to Christel Vrain (christel.vrain@univ-orleans.fr)

- CV
- One-page cover letter (clearly indicating available start 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
- **Attention:** all documents should be in English or French.