Clustering collaboratif : le challenge de regrouper conjointement

Germain Forestier

INRIA / INSERM / CNRS / Univ. Rennes 1, VISAGES U746, Rennes, France
germain.forestier@inria.fr

Résumé. Le clustering collaboratif consiste à faire collaborer conjointement plusieurs méthodes de clustering afin de parvenir à un résultat consensuel. Les différentes méthodes impliquées dans la collaboration vont partager des informations et vont remettre en cause leurs décisions en fonction des solutions proposées par les autres méthodes. Un enjeu important consiste à faire collaborer des méthodes différentes tout en assurant une prise en compte des décisions de chacune d'entre elles. Cet article présente les principales approches en clustering collaboratif et contextualise nos travaux dans ce domaine. Enfin, quelques enjeux émergents sont également exposés.

1 Introduction

L'idée de combiner les décisions de plusieurs méthodes de classification non supervisée a émergé du travail important mené dans le domaine de la combinaison de méthodes supervisées (Kuncheva, 2008). Dans le cas supervisé, le travail est simplifié par l'existence d'une référence commune (*i.e.* les classes) qui peut servir à faciliter la combinaison des décisions.

A contrario, dans le cadre de la combinaison de méthodes non supervisées, il n'existe pas de liens évident entre les clusters des différents résultats et les ceux-ci n'ont pas forcément le même nombre de clusters. D'autres approches ont donc été envisagées pour permettre la combinaison de ces résultats hétérogènes. De plus, d'autres contraintes comme la distribution des données sur plusieurs sites, ou le respect de la confidentialité des données, rendent parfois impossible un traitement centralisé et constituent une autre motivation de ces approches.

Afin d'adresser ce problème, nous nous sommes intéressé au clustering collaboratif qui consiste à faire collaborer plusieurs méthodes de clustering conjointement. Ces différentes méthodes vont partager des informations et vont remettre en cause leurs décisions en fonction des décisions proposées par les autres méthodes. Ainsi, une discussion est entreprise entre les méthodes afin de faire converger collectivement les différents résultats.

Dans cet article, nous présentons les principales approches en clustering collaboratif afin de contextualiser nos travaux et d'illustrer les résultats obtenus. Enfin, les enjeux émergents de ce domaine sont abordés.

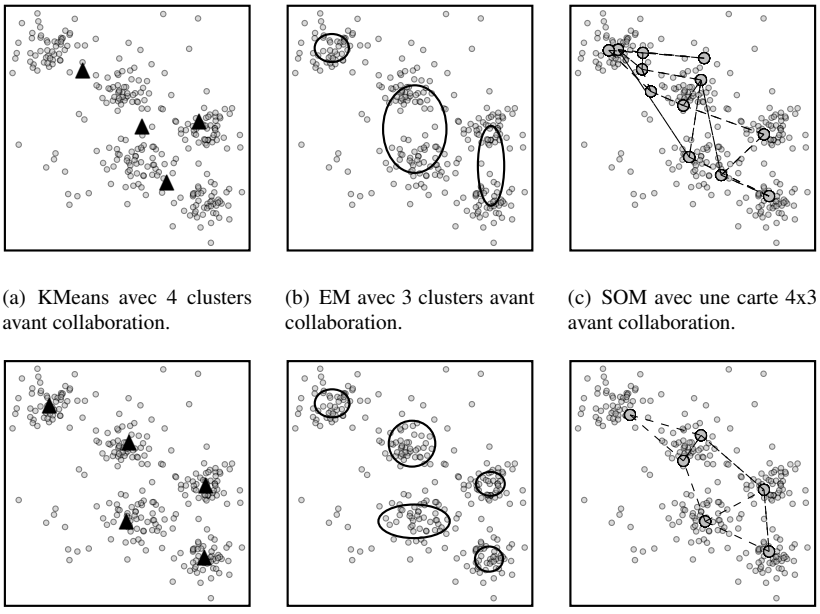
2 Le clustering collaboratif

Plusieurs approches existent en clustering collaboratif mettant en œvre la collaboration à différents niveaux. Les méthodes dîtes de clustering par ensemble (Strehl et Ghosh, 2002), s'intéressent à la combinaison de résultats de clustering en étudiant uniquement l'affectation des objets aux clusters des résultats. De ce fait, les données utilisées pour générer les résultats ne sont plus utilisées lors du processus collaboratif, qui consiste alors à trouver une partition consensuelle résumant l'ensemble des partitions de départ. Ces approches font ainsi l'hypothèse que les partitions générées initialement sont de qualité suffisante pour permettre au processus collaboratif de trouver un résultat pertinent. Les approches de clustering multi-objectifs (Handl et Knowles, 2007) tentent de réduire cette hypothèse, en créant de nouvelles partitions à partir des partitions initiales à l'aide d'opérateurs de croisement et de mutation. Cependant, ces opérateurs ne mettent pas en œuvre les méthodes de clustering utilisées pour produire les partitions initiales, introduisant de ce fait un nouveau biais.

Une autre approche en clustering collaboratif, consiste à se concentrer sur une méthode de clustering, en étudiant la manière dont peuvent collaborer plusieurs modèles construits par cette méthode. Dans ce cadre, les travaux de Pedrycz (2007) se sont intéressés à l'algorithme Fuzzy-c-means et proposent une méthode consistant à comparer les degrés d'appartenance des objets aux centroides des différents résultats. De manière similaire, Cleuziou et al. (2009) se sont également intéressés à l'algorithme Fuzzy-c-means et ont proposé une méthode de clustering collaboratif en introduisant un terme de pénalité permettant d'évaluer et de réduire itérativement les désaccords entre les résultats. Une autre approche, proposée par Grozavu et Bennani (2010), s'est intéressée à l'algorithme des cartes de Kohonen (SOM) et à la combinaison de plusieurs cartes.

Ces différentes approches permettent d'optimiser la collaboration pour une méthode donnée et ont obtenu de très bons résultats pour le clustering collaboratif. Cependant, elles restent spécifiques à l'algorithme de clustering utilisé et sont difficilement généralisables. Dans la quête d'une méthode plus générique, permettant l'utilisation conjointe de différents algorithmes de clustering, nous nous sommes intéressés dans nos travaux à la réduction des désaccords entre résultats de clustering provenant de méthodes de clustering différentes. Ce choix pose la question importante de la possibilité de partager des informations entre des résultats dont la structure sous-jacente des modèles n'est pas similaire. De ce fait, il est nécessaire de proposer un mécanisme générique, permettant d'une part d'identifier les désaccords, et d'autre part d'effectuer des actions permettant leur résolution.

Dans le cadre de nos travaux (Forestier, 2010), nous avons proposé une approche de clustering collaboratif permettant la collaboration de résultats produits par des méthodes différentes. Les désaccords entre les résultats proposés par les différentes méthodes sont appelés des *conflits*. Ces conflits sont identifiés en observant la répartition des objets dans les clusters des différents résultats. Cette étape d'identification est donc indépendante des méthodes de clustering utilisées. A l'issue de cette étape d'identification, une tentative de résolutions des conflits est engagée. Cette étape de résolution consiste à appliquer des opérateurs (*e.g.* fusion de clusters, scission de clusters) aux résultats impliqués dans le conflit. L'application de ces opérateurs est dépendante de la méthode de clustering utilisée, permettant ainsi de mener des modifications locales à partir d'une information obtenue de manière globale. La seule contrainte concernant la participation d'une méthode de clustering à la collaboration, est sa capacité à implémenter ces opérateurs. La résolution des conflits peut être itérative, ou faire intervenir des métaheuris-



(a) KMeans avec 4 clusters avant collaboration.

(b) EM avec 3 clusters avant collaboration.

(c) SOM avec une carte 4x3 avant collaboration.

(d) KMeans avec 5 clusters après collaboration.

(e) EM avec 5 clusters après collaboration.

(f) SOM avec 6 neurones après collaboration.

FIG. 1 – *Résultats de clustering avant (a,b,c) et après (d,e,f) collaboration pour trois résultats produits avec trois méthodes différentes (KMeans, EM et SOM).*

tiques (Forestier et al., 2010) plus complexes pour mieux parcourir l'espace de recherche des solutions. La figure 1 présente un exemple de collaboration entre trois résultats de clustering obtenus avec trois méthodes différentes : KMeans, EM et SOM. Elle illustre les états respectifs des résultats avant et après collaboration pour chacune des méthodes. La forte similarité des résultats obtenus après collaboration, malgré la diversité des méthodes mises en œuvre, atteste de la capacité de notre approche à faire collaborer des méthodes de clustering différentes.

3 Discussion

Le clustering collaboratif est un domaine récent dont les fondements commencent à être posés. La multiplication des travaux s'intéressant à la combinaison et à l'intégration de plusieurs résultats de clustering montrent l'intérêt de la communauté scientifique pour ces approches.

Dans le cadre de nos travaux, nous avons proposé une méthode collaborative permettant l'échange d'information entre des résultats de clustering produits par des méthodes différentes. Cependant, elle n'adresse pas directement le problème de l'apprentissage collaboratif. En effet, un des prochains challenges à résoudre consistera à échanger des informations directement pendant l'étape d'apprentissage des modèles (calcul des centroides pour KMeans, création de la carte pour SOM, etc.). Il conviendra alors de développer des processus génériques permettant l'échange d'informations entre les méthodes.

Enfin, un autre défi concerne l'utilisation de méthodes de clustering collaboratif sur plusieurs sources de données fortement hétérogènes. En effet, bien que certaines méthodes s'intéressent au traitement de données multivues (données représentées par différents ensembles d'attributs), ces données ont toujours le même révérenciel, c'est-à-dire le même nombre d'objets. Ainsi, un autre enjeu consistera à étudier comment différentes vues plongées dans des référentiels différents, pourraient être utilisées conjointement dans un processus collaboratif. Nos travaux (Wemmert et al., 2009) ont proposé des solutions dans le cadre du traitement collaboratif d'images de télédétection à différentes résolutions (*i.e.* taille des images différentes), nous poussant ainsi à étudier la définition de critères de *cohérence* entre résultats de clustering. Des critères génériques à tout type de données sont encore à proposer.

Références

Cleuziou, G., M. Exbrayat, L. Martin, et J.-H. Sublemontier (2009). CoFKM : A centralized method for multiple-view clustering. In *IEEE International Conference on Data Mining*, pp. 752–757.

Forestier, G. (2010). *Connaissances et clustering collaboratif d'objets complexes multisources*. Ph. D. thesis, Université de Strasbourg.

Forestier, G., C. Wemmert, et P. Gancarski (2010). Towards conflict resolution in collaborative clustering. In *IEEE International Conference on Intelligent Systems*, pp. 361–366.

Grozavu, N. et Y. Bennani (2010). Classification collaborative non supervisée. In *Conférence francophone sur l'apprentissage automatique (CAP)*.

Handl, J. et J. Knowles (2007). An evolutionary approach to multiobjective clustering. *IEEE Transactions on Evolutionary Computation 11*(1), 56–76.

Kuncheva, L. I. (2008). Classifier ensembles : Facts, fiction, faults and future. In *International Conference on Pattern Recognition, Plenary talk*.

Pedrycz, W. (2007). Collaborative and knowledge-based fuzzy clustering. *International Journal of Innovative, Computing, Information and Control 1*(3), 1–12.

Strehl, A. et J. Ghosh (2002). Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research 3*, 583–617.

Wemmert, C., A. Puissant, G. Forestier, et P. Gancarski (2009). Multiresolution remote sensing image clustering. *IEEE Geoscience and Remote Sensing Letters 6*, 533 – 537.

Summary

Collaborative clustering consists to make jointly collaborate several clustering methods in order to improve their results. The different methods involved in the collaboration share their informations and question their decisions according to the decisions proposed by the other methods. An important challenge is to make collaborate different clustering methods while insuring that each point of view is taken into consideration. This article present the main approaches in collaborative clustering and present our work in that field. Furthermore, some emerging challenges are also presented.