Un Jeu de Potentiel pour le Clustering de Données

K. BOUCHAMA, M.S. RADJEF, L. SAIS, A. LALLOUET

Doctoriales de Recherche Opérationnelle, Le 12 et 13 Décembre 2018

Introduction et Position du Problème

Actuellement, les données ont un intérêt considérable et croissant dans les secteurs économiques et industriels. De nouvelles techniques sont apparues pour gérer ce flux important de données qui peuvent être de différentes sources et de différents types. L'une des techniques les plus répandues est le clustering.

Le clustering est une classification non supervisée de données sur la base de leur similarité. Bien qu'il existe plusieurs méthodes pour la résolution de ce problème, aucune approche reconnue fiable ne peut garantir la qualité des solutions fournies. Chaque méthode présente des avantages et des inconvénients selon le(s) critère(s) d'évaluation utilisé(s).

Depuis peu, le clustering constitue un nouveau champ d'application de la théorie des jeux. Le tout premier travail initiant cette application est un article de *M. Pellilo et al,* intitulé : *Grouping with asymmetric affinities: A game-theoretic perspective[1]*. L'intérêt d'aborder le problème de clustering par les outils de la théorie des jeux est qu'ils permettent de surpasser quelques limitations des approches par partitionnement, tels que la connaissance au préalable du nombre de groupes à former, le chevauchement des clusters, la détection des bruits, la symétrie des matrices de similarité, etc.

Notre principal objectif dans ce travail est alors de proposer une nouvelle approche de clustering, basée sur les outils de la théorie des jeux, qui permettrait de ramener la résolution du problème de clustering de données à la résolution d'un jeu de potentiel, dont l'existence d'au moins un équilibre de Nash en stratégies pures est garantie. Pour résoudre ce jeu, nous proposons un algorithme efficace, basé sur la notion de stratégies de meilleure réponse pour la recherche d'un équilibre de Nash en stratégies pures, qui dans ce cas, correspond à une solution stable au problème de clustering.

Modélisation du Problème

Avant de présenter le modèle de jeu que nous proposons pour le problème de clustering, nous allons considérer les éléments suivants:

- $\mathcal{O} = \{o_1, \ldots, o_n\}$, $o_i \in \mathbb{R}^r, i = \overline{1,n}$ un ensemble de n objets;
- $\{C_1, \ldots, C_k\}$, l'ensemble des clusters à former à partir de $\mathcal{O}$;
- $\{c_1, \ldots, c_k\}$, leurs centroids.

A partir de là, nous pouvons construire le jeu noncoopératif associé à ce problème de clustering de données comme suit:

- Associons à chaque objet $o_i \in \mathcal{O}$, un joueur $i \in P$. $P = \{1, \ldots, n\}$ est l'ensemble des joueurs.
- Considérons que l'ensemble des strategies d'un joueur $i \in P$ coincide avec l'ensemble des clusters i.e $S_i = \{C_1, \ldots, C_k\}, \forall i \in P$.
- La fonction d'utilité que chaque joueur vise à minimiser est donnée par:

$$U_i^p(s_i, s_{-i}) = d(o_i, c_{s_i})^2 \times \alpha$$

Avec α est un paramètre qui dépend de s_i

Le jeu correspondant au problème de clustering est alors défini par:

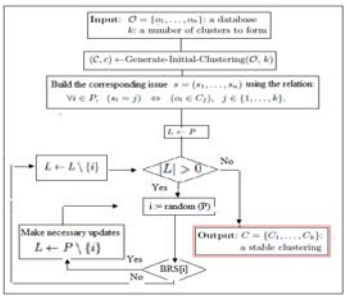
$$G^p = \langle P, \{S_i\}_{i \in P}, \{U_i^p\}_{i \in P} \rangle$$

Quelques Résultats Théoriques

- Le jeu de clustering G^p est un jeu de potentiel ;
- Le jeu de clustering G^p admet au moins un PNE.
- Tout équilibre de Nash pure $s^* \in S$ du jeu G^p correspond à une solution de clustering stable.

Résolution du Jeu de Clustering

Rechercher une solution stable au problème de clustering revient à rechercher un équilibre de Nash pure pour le jeu de potentiel correspondant. Pour cela, nous proposons un algorithme itératifs basé sur les dynamiques de recherche des stratégies de meilleures réponses, nommé « *IBRC* ». Cet algorithme nécessite de calculer une solution initiale, en affectant n objets à k clusters dont les centroides seront générés aléatoirement. Nous obtenons un premier clustering sur lequel nous allons construire notre modèle de jeu. L'algorithme IBRC est résumé par l'organigramme suivant:



Résultats Expérimentaux

L'algorithme proposé a été implémenté puis testé sur une série de bases de données résumée dans le tableau suivant:

Datasets dimensions		n	r	k
Synthetic datasets	R15	600	2	15
	D31	3100	2	31
	Dim32	1024	32	16
	Dim64	1024	64	16
	Dim128	1024	128	16
	Dim256	1024	256	16
	Dim512	1024	512	16
	Dim1024	1024	1024	16
	Unbalance	6500	2	8
Real datasets	Iris	150	4	3
	Ecoli	336	8	8
	Glass	214	9	7
	Thyroid	215	5	2
	Mushroom	8124	23	2

Les résultats obtenues par notre algorithme ont été comparés à ceux fournis par l'algorithme classique du k-means. De plus, nous avons proposé une version hybride d'IBRC, nommée *H-IBRC,* où nous considérons que la solution initiale de l'algorithme IBRC est une solution fournie par l'algorithme du k-means.

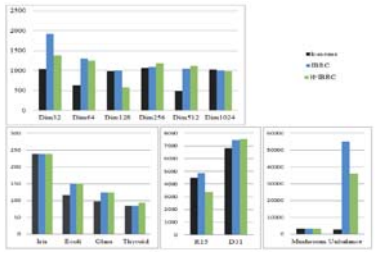
Les solutions fournies par chacun de ces trois algorithmes ont été évaluées sur plusieurs critères (internes et externes).

➢ L'index de Calinski-Harabasz (CH-index):

Cette mesure permet d'évaluer la validité des clusters formés. Sa formulation mathématique est la suivante:

$$CH\text{-}index = \left(\sum_{i=1}^{k} |C_i| d^2(c_i, c)/(k-1) \right) \Big/ \left(\sum_{i=1}^{k} \sum_{x \in C_i} d^2(x, c_i)/(n-k) \right)$$

Les résultats obtenus sont résumés dans la figure suivante. Nous remarquons clairement que d'une part, les valeurs du CH-index pour les solutions fournies par l'algorithme IBRC sont significativement meilleures comparées à celles fournies par k-means sur la plus part des instances testées. D'autre part, l'algorithme H-IBRC permet d'améliorer les valeurs de CH-index retournées par le k-means.



Evaluation des solutions par leur valeur du CH-index

➢ La Pureté:

La mesure de pureté permet de mesurer le nombre moyens d'objets correctement affectés. Sa formule est la suivante:

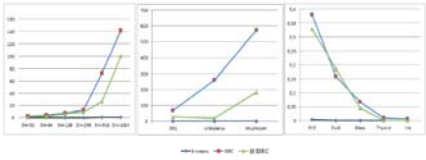
$$Purity = \sum_{j=1}^{k} \frac{|C_j|}{n} P_j \qquad P_j = \frac{1}{|C_j|} \max_i (n_j^i), \ j = \{1, \ldots, k\}$$

Les résultats obtenus sont résumés dans le tableau suivant:

	k-means	IBRC	H-IBRC
Dim32	0.8837	0.8876	0.8906
Dim64	0.9062	0.8926	0.8857
Dim128	0.9111	0.8867	0.8984
Dim256	0.8994	0.8916	0.8896
Dim512	0.9003	0.8896	0.8916
Dim1024	0.9023	0.8906	0.8867
R15	0.8133	0.8233	0.8233
Ecoli	0.6845	0.6607	0.6607
Glass	0.7617	0.7242	0.7289
D31	0.7787	0.8093	0.8125
Mushroom	0.9643	0.9612	0.9612
Unbalance	0.7821	0.8060	0.7824
Thyroid	0.9120	0.8980	0.898
Iris	0.9000	0.8530	0.853

Pour cette mesure, nous constatons que les meilleurs résultats sont fournis par k-means, Cependant, la différence moyenne avec la pureté des clusters formés par IBRC et H-IBRC ne dépasse pas 1,5%,

➢ Analyse des Temps CPU:



Cette figure révèle une grande vitesse de convergence du k-means, contrairement au temps CPU d'IBRC qui s'accroit exponentiellement en fonction de la dimension de l'instance testée. Par contre, H-IBRC tire bien profite de la vitesse du k-means pour réduire considérablement le temps de calcul de l'algorithme IBRC.

Conclusion

Dans ce travail, nous avons proposé un jeu noncoopératif admettant une fonction de potentiel, pour représenter le problème de clustering de données. Après avoir démontré que tout équilibre de Nash pure de ce jeu correspond à un clustering stable, nous avons proposé un algorithme IBRC basé sur la recherche itératives des stratégies de meilleures réponses, que nous avons par la suite hybridé avec l'algorithme du k-means. Les tests effectués sur ces algorithmes ont témoignés de la bonne qualité des solutions fournies par nos algorithmes. De plus, l'hybridation H-IBRC est très efficace pour la réduction du temps de convergence d'IBRC.

Références

1. Torsello, A., Bulo, S. R., & Pelillo, M. (2006, June). Grouping with asymmetric affinities: A game-theoretic perspective. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (Vol. 1, pp. 292-299). IEEE.
2. Feldman, M., Lewin-Eytan, L., and Seffi Naor, J., Hedonic clustering games. In Proceedings of ACM symposium on Parallelism in algorithms and architectures, pages 267–276. ACM, 2012.
3. Marden, J., State based potential games. Automatica, 48(12): 3075–3088, 2012.