Subspace Co-clustering avec Convolution Bilatérale sur Graphe

Chakib Fettal[*,**], Lazhar Labiod[*], Mohamed Nadif[*]

[*] Centre Borelli, UMR 9010
Université Paris Cité
{prenom.nom}@u-paris.fr
[**]Informatique Caisse des Dépôts et Consignations

Résumé. Le *subspace clustering* vise à partitionner un ensemble d'observations de haute dimension. Si cette approche a donné de bons résultats dans le domaine du partitionnement d'images, elle s'est avérée inefficace pour le partionnement de données sparses comme c'est le cas des données matrices termes-documents. Une extension appropriée de cette approche au co-clustering, particulièrement efficace sur des données sparses, s'avère utile pour traiter de données attribuées. Ainsi, nous traitons le problème de la sparsité par le biais d'une convolution bilatérale sur graphe qui favorise l'effet de regroupement. Nous montrons la compétitivité de notre modèle par rapport à l'état de l'art sur des ensembles de données de graphes attribués en termes de performance et d'efficacité computationelle.

1 Introduction et Contexte

Le présent article est un résumé de l'article publié dans la conférence CIKM (Fettal et al., 2022b). Le subspace clustering (Parsons et al., 2004) consiste à regrouper des observations (ou éléments) en fonction des sous-espaces qui les contiennent. Il existe une variété d'approches pour résoudre ce problème, dont beaucoup d'entre elles considèrent la formulation auto-expressive où l'on suppose que chaque élément peut être écrit comme une combinaison linéaire des éléments dans le même sous-espace. Typiquement, la formulation générique est donnée par

$$\min_{\mathbf{R}} \quad \| \mathbf{X} - \mathbf{R}\mathbf{X} \|^2 + \Omega(\mathbf{R}) \quad \text{s.t.} \quad \mathbf{R} \in \mathcal{R} \tag{1}$$

où $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{R} \in \mathbb{R}^{n \times n}$ est appelée la matrice d'auto-représentation, $\Omega(\mathbf{R})$ sert de terme de régularisation pour induire des propriétés souhaitables sur $\mathbf{R}$ et éviter les solutions triviales (telle que $\mathbf{R} = \mathbf{I}$), et $\mathcal{R}$ est la région réalisable.

Etant donné une solution optimale $\mathbf{R}^*$, une matrice d'affinité est générée sur la base des amplitudes des entrées de $\mathbf{R}^*$, en utilisant généralement $|\mathbf{R}^* + \mathbf{R}^{*\top}|/2$, et une partition des observations est ensuite générée en utilisant une méthode de partitionnement de graphes, par exemple l'algorithme de partitionnement spectral (Shi et Malik, 2000).

Les méthodes de type subspace clustering basées sur la propriété d'auto-expression (Zhang et al., 2021) ont été largement utilisées pour regrouper des ensembles de données de type image