Méthode ensemble de clustering profond

Séverine Affeldt*, Lazhar Labiod*, Mohamed Nadif*

*Université de Paris, CNRS, Centre Borelli, F-75006 Paris
<prénom.nom>@u-paris.fr

Résumé. Plusieurs stratégies de classification non supervisée combinant des algorithmes de *clustering* classiques et les auto-encodeurs ont été étudiées. Ces stratégies améliorent généralement les performances de clustering, mais dépendent des données et des hyperparamètres choisis qu'on ne peut fixer en amont pour un apprentissage non supervisé. Pour atténuer l'impact de la configuration des hyperparamètres sur la qualité du partitionnement, nous proposons une approche *ensemble*, ne nécessitant aucun pré-entraînement, qui exploite le clustering spectral et les forces de l'auto-encodeur. Notre approche génère plusieurs encodages profonds à partir des données d'origine, construit une matrice d'affinité *consensus* clairsemée et de faible dimension via la stratégie des *anchors* puis applique la classification spectrale pour obtenir l'espace commun partagé par les encodages. La stratégie des *anchors* assure une représentation efficace des encodages et leur fusion permet d'atténuer les problèmes d'hyperparamètres (Affeldt et al., 2020).

1 Introduction

Malgré leur succès, la plupart des méthodes de clustering sont mises à mal par les données générées avec les applications actuelles, qui sont généralement de grande dimension, hétérogènes et clairsemées. De nombreux chercheurs ont étudié de nouveaux modèles de clustering pour surmonter ces difficultés. Une catégorie prometteuse de ces modèles repose sur l'*embedding* des points (Allab et al., 2016, 2018; Labiod et Nadif, 2020). Dans l'apprentissage profond, les auto-encodeurs (DAE : Deep AutoEncoders) se sont avérés également utiles dans cette tâche (Hinton et Zemel, 1994); les auto-encodeurs peuvent transformer des données de manière non linéaire dans un espace latent. Un auto-encodeur se compose généralement d'un *encoder*, qui peut fournir un encodage des données d'origine en dimension inférieure, et d'un *decoder*, pour définir le coût de reconstruction des données. Dans le contexte du clustering, l'idée générale est d'intégrer les données dans un espace latent de faible dimension, puis d'effectuer le clustering dans ce nouvel espace.

Plusieurs travaux intéressants ont récemment combiné l'apprentissage profond de représentations et le clustering. Ces deux tâches peuvent être associées de manière séquentielle ou conjointe. Ainsi, Tian et al. (2014) utilisent un auto-encodeur *empilé* pour apprendre une représentation du graphe d'affinité, puis exécutent *k*-means sur les représentations apprises pour obtenir les clusters. Le processus d'embedding ne garantissant pas une représentation adaptée