Clustering Visuel Semi-Supervisé pour des systèmes en *coordonnées en étoiles* 3D

Loïc Lecerf*, Boris Chidlovskii*

*Xerox Research Centre Europe
6, chemin de Maupertuis,
38240 Meylan, France
{Prenom.Nom}@xrce.xerox.com,

Résumé. Dans cet article, nous proposons une approche qui combine les méthodes statistiques avancées et la flexibilité des approches interactives manuelles en clustering visuel. Nous présentons l'interface *Semi-Supervised Visual Clustering* (SSVC). Sa contribution principale est l'apprentissage d'une métrique de projection optimale pour la visualisation en *coordonnées en étoiles* ainsi que pour l'extension 3D que nous avons développée. La métrique de distance de projection est apprise à partir des retours de l'utilisateur soit en termes de similarité/dissimilarité entre les items, soit par l'annotation directe. L'interface SSVC permet, de plus, une utilisation hybride dans laquelle un ensemble de paramètres sont manuellement fixés par l'utilisateur tandis que les autres paramètres sont déterminés par un algorithme de distance optimale.

1 Introduction

Obtenir un clustering efficace et de haute qualité sur des données de grande taille est un problème majeur pour l'extraction des connaissances. Il existe une demande de plus en plus importante pour des techniques flexibles et efficaces de clustering capables de s'adapter à des jeux de données de structure complexe. Un ensemble de données est typiquement représenté dans un tableau composé de N items (lignes) et d dimensions (colonnes). Un item représente un événement ou une observation, alors qu'une dimension peut-être un attribut ou une caractéristique de l'item. Dans un mode *semi-supervisé* ou *supervisé*, une partie ou tous les items peuvent être annotés par une classe. Les méthodes de clustering tentent de partitionner les items en groupes avec une mesure de similarité. Un ensemble de données peut être grand en termes de nombre de dimensions, nombre d'éléments, ou les deux.

L'approche classique est basée sur des algorithmes de clustering, comme les K-moyennes, le clustering spectral ou hiérarchique ainsi que leurs multiples variantes (Hastie et al., 2001). Il existe cependant plusieurs inconvénients connus à ces méthodes. Premièrement, il n'est pas toujours facile de déterminer, visualiser et valider les clusters de forme irrégulière. Plusieurs algorithmes sont efficaces pour trouver des clusters dans des formes elliptiques (donc convenant aux distributions normales multidimensionnelles), mais peuvent échouer à reconnaître des clusters de forme complexe. Deuxièmement, les algorithmes existant sont automatiques, ils excluent toute intervention de l'utilisateur dans le processus jusqu'à la fin de l'algorithme.