Une bibliothèque de *subspace clustering*

Arthur Guillon, Marie-Jeanne Lesot & Christophe Marsala

1 Sujet : le *subspace clustering*

Le *clustering* (partitionnement) est une tâche d'apprentissage non-supervisé qui consiste à circonscrire au sein d'un jeu de données des groupes de données fortement similaires entre elles, et fortement dissimilaires des autres données. Le *subspace clustering* [4, 5] est une généralisation de cette tâche, dans laquelle les groupes sont identifiés en même temps que les sous-espaces dans lesquels ils se trouvent.

Selon le contexte d'application, différents types de sous-espaces peuvent être considérés : simple sélection ou pondération d'attributs, recherche de sous-espaces linéaires non-parallèles aux axes… Des exemples canoniques sont les algorithmes CLIQUE [1], SSC [2] ou WFCM [3].

2 Objet : une plate-forme logicielle de subspace clustering

Le projet consiste en l'implémentation dans le langage Python d'une bibliothèque regroupant plusieurs algorithmes de subspace clustering, qui visera à faciliter l'expérimentation et la comparaison des performances de ces algorithmes. Par conséquent, la bibliothèque devra :

- implémenter quelques-uns des algorithmes de subspace clustering les plus connus ;
- faciliter la visualisation des résultats des algorithmes ;
- fournir des primitives de génération de données artificielles ;
- faciliter l'expérimentation et la comparaison des algorithmes implémentés sur des données artificielles et réelles.

Un aspect central du projet est la conception d'une architecture générique et extensible, réutilisable pour l'implémentation de nouveaux algorithmes de subspace clustering.

Références

[1] Agrawal, R., Gehrke, J., Gunopulos, D. & Raghavan, P. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. In *Proc. of the 1998 ACM SIGMOD Int. Conf. on Management of Data*. Seattle, Washington, USA : ACM, 1998, p. 94–105.

[2] Elhamifar, E. & Vidal, R. Sparse subspace clustering. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. 2009, p. 2790–2797.

[3] Keller, A. & Klawonn, F. Fuzzy clustering with weighting of data variables. In *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 8.06 (2000), p. 735–746.

[4] Kriegel, H.-P., Kröger, P. & Zimek, A. Clustering high-dimensional data : A survey on subspace clustering, pattern-based clustering, and correlation clustering. In *ACM Transactions on Knowledge Discovery from Data (TKDD)* 3.1 (2009).

[5] Vidal, R. A tutorial on subspace clustering. In *IEEE Signal Processing Magazine* 28.2 (2010), p. 52–68.