Clustering avec R
Application sur les données des eaux embouteillées commercialisées en Algérie

Y. Khemal-Bencheikh [a] & I. Naas [b]

[ab] Laboratoire de Mathématiques Fondamentales et Numériques LMFN
Département de Mathématiques-Faculté des Sciences
UFA Sétif1 El-Bez 19000. Algérie

bencheikh-00@univ-setif.dz [a]
sherry.ihcene19@gmail.com [b]

Mots clefs : Clustering, K-means, analyse en composantes principales, clustering hiérarchique

Le regroupement d'objets, dans un cadre non-supervisé est une tâche importante et difficile en apprentissage. Ce processus qu'on appelle clustering [1] intervient dans des contextes variés tels que la découverte des connaissances, la simplification dans la représentation ou la description d'un ensemble de données. Malgré le nombre important de méthodes existantes; plusieurs problématiques restent encore ouvertes dans le cadre du clustering : comme la difficulté de fixer les paramètres d'entrés par l'utilisateur ou une représentation des résultats sous forme facilement interprétable ou encore l'évaluation des résultats et la comparaison des différentes méthodes.

Nous proposons dans ce papier, d'utiliser les méthodes du clustering pour partitionner 20 marques d'eaux embouteillées commercialisées en Algérie. Le tableau de données sera traité à l'aide d'un des algorithmes du clustering par partitionnement : le clustering par K-means. Nous comparons les résultats obtenus avec d'autres techniques d'analyse de données à savoir l'analyse en composantes principales et le clustering hiérarchique. Cette étude nous a permis de donner des indications concernant les eaux minérales naturelles et les eaux de sources en Algérie et les répartir en plusieurs clusters suivant leurs caractéristiques physico-chimiques. Ces indications peuvent être utiles pour le consommateur et pour des questions économiques.

Le traitement des données a été réalisé en utilisant le logiciel R et le package ade4TkGUI [2], qui propose une interface utilisateur graphique pour les fonctions de base du package d'analyse statistique de données multi variées ade4.

Références

[1] Ben-Dor, A., Shamir, R., Yakhin, Z. (1990). Clustering gene expression patterns . Journal of computational biology 6 (3/4), 281-297.

[2] Herrington R (2006). ade4TkGUI - A GUI for Multivariate Analysis and Graphical Display in R. Benchmarks Online, **9**(12).