Clustering divisif monothétique. Le package divclust

M. Chavent[a,b] and **M. Fuentes**[b]

[a]Université de Bordeaux,
351 cours de la libération, 33405 Talence,
marie.chavent@u-bordeaux.fr

[b]Inria Bordeaux Sud-Ouest,
200 Avenue de la Vieille Tour, 33405 Talence
marc.fuentes@inria.fr

Mots clefs : Clustering descendant hiérarchique, dendrogramme de décision, données mixtes.

DIVCLUS-T [2] est une méthode descendante de clustering hiérarchique basée sur une approche de bi-partitionnement monothétique permettant de lire le dendrogramme comme un arbre de décision [1]. Comme pour la méthode ascendante de Ward, le critère de qualité d'une partition est l'inertie intra-classe. A chaque étape de l'algorithme, une classe est divisée en deux selon une question binaire (définie sur une seule variable), mais on choisit dans l'ensemble de toutes les questions binaires possibles celle qui induit la partition de plus petite inertie intra-classe. Cette inertie intra-classe quant à elle est calculée de manière classique sur toutes les variables. Ainsi si n observations sont décrites par p variables dans un noeud par exemple, il y aura $p \times (n-1)$ questions binaires possibles et donc $p \times (n-1)$ bi-partitions des n observations à évaluer. La bi-partition ayant l'inertie intra-classe la plus petite est alors retenue et la question binaire caractérise cette division. La méthode DIVCLUS-T construit ainsi une hiérarchie indicée comme celle de Ward, mais le dendrogramme fournit pour chaque classe une règle de décision, condition nécessaire et suffisante d'appartenance à la classe (voir Figure 1).

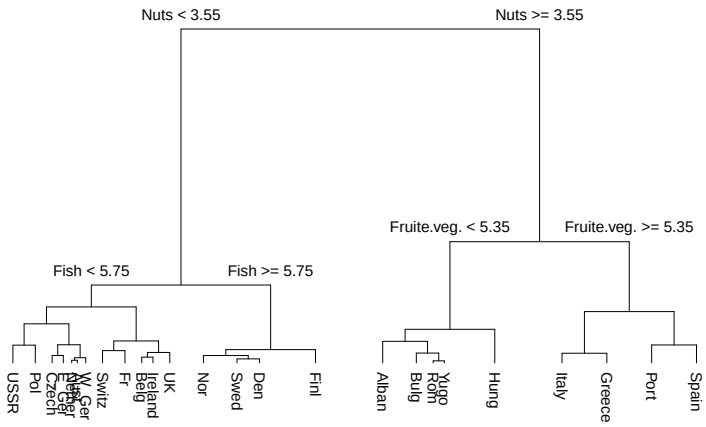

Figure 1: Dendrogram de DIVCLUS-T pour les données de consommation en protéines dans 25 pays européens.

Ce qui caratérisique également l'algorithme est le fait que les classes ne sont pas systématiquement divisées jusqu'à l'obtention des singletons. En effet, à chaque étape, c'est la classe qui

induit la plus grand variation d'inertie qui est choisie pour être divisée et les divisions peuvent s'arrêter après $K-1$ étapes, le nombre K étant donné en entrée par l'utilisateur. La dernière partition obtenue est alors la partition en K classes. Ainsi, le dendrogramme obtenu représente uniquement les partitions de 2 à $K-1$ classes correspondant à la partie "haute" du dendrogramme complet (dendrogramme qui descend jusqu'aux singletons). Par exemple sur la Figure 2, les divisions ont étés arrêtées après 4 étapes et le dendrogramme est exactement le haut du dendrogramme de la Figure 2 .

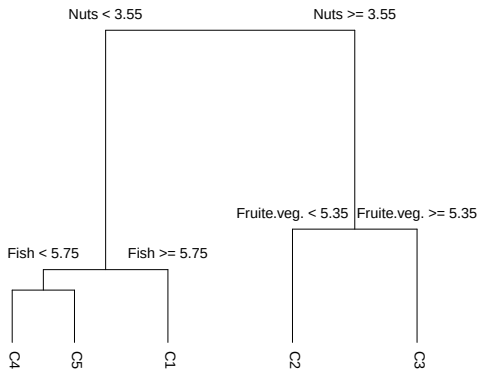

Figure 2: Dendrogramme de DIVCLUS-T correspondant aux partitions de 2 à 5 classes pour les données de consommation en protéines dans 25 pays européens.

Nous verrons enfin comment cette méthode peut s'appliquer à des données quantitatives, qualitatives ou mixtes [3]. L'algorithme ainsi que le package **divclust** seront présentés et illustrés sur des exemples.

Références

[1] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). Classification and regression Trees. C.A:Wadsworth.

[2] Chavent, M., Briant O. and Lechevallier, Y. (2007). DIVCLUS-T: a monothetic divisive hierarchical clustering method. . *Computational Statistics and Data Analysis*, **32**, 687-701.

[3] Chavent, M., Kuentz, V., Labenne, A., and Saracco, J., Multivariate analysis of mixed data: The PCAmixdata R package, arXiv:1411.4911 [stat.CO]