Classification divisive sous contraintes. Application au bassin de la Charente

Marie Chavent, Yves Lechevallier, Kevin Petit, Françoise Vernier

Classification divisive sous contraintes. Application au bassin de la Charente

Marie Chavent & Yves Lechevallier & Kevin Petit & Francoise Vernier

Marie Chavent
Université Bordeaux 2, Institut de Mathématiques de Bordeaux, UMR 5251.
146, Rue Léo Saignat, 33076 Bordeaux cedex, France
Yves Lechevallier
INRIA, Paris-Rocquencourt 78153 Le Chesnay cedex, France
Kevin Petit et Francoise Vernier
CEMAGREF-Bordeaux, Unité de recherche ADER 50,
Avenue de Verdun, 33612 Cestas, France

Résumé : DIVCLUS-T est une méthode de classification hiérarchique descendante et monothétique qui cherche à optimiser à chaque étape un critère d'homogénéité défini sur la partition construite. Le dendrogramme a la particularité d'être expliqué, à chaque palier, par une question binaire. Nous proposons dans ce papier une nouvelle version de cette méthode appelée C-DIVCLUS-T qui est capable de prendre en compte les contraintes de contiguïté. Nous appliquons C-DIVCLUS-T sur un ensemble zones hydrologiques qui sont décrites par des variables agricoles et environnementales.

Abstract: DIVCLUS-T is a descendant hierarchical clustering algorithm based on a monothetic bipartitional approach allowing the dendrogram of the hierarchy to be read as a decision tree. We propose in this paper a new version of this method called C-DIVCLUS-T which is able to take contiguity constraints into account. We apply C-DIVCLUS-T to hydrological areas described by agricultural and environmental variables, in order to take their geographical contiguity into account in the monothetic clustering process.

1 Introduction

La méthode DIV [4] est une méthode de classification descendante hiérarchique qui s'applique sur l'ensemble des objects à classer, noté Ω , par divisions successives et s'articule autour des trois points suivants:

- Les divisions s'arrêtent après K étapes. On obtient donc le "haut" du dendrogramme c'est à dire les partitions de 2 à $K + 1$ classes.
- A chaque étape cette méthode choisit de diviser la classe telle que la nouvelle partition ainsi obtenue maximise le gain entre le critère d'homogénéité de la classe divisée et la somme des critères d'homogénéité des deux classes issues de cette division.

- Le principe de DIV dans l'étape de bi-partitionnement d'une classe à n éléments en deux sous-classes est d'éviter d'évaluer l'homogénéité des $2^{n-1}-1$ bi-partitions possibles pour en retenir la meilleure, mais d'évaluer uniquement ce critère d'homogénéité sur l'ensemble de toutes les bi-partitions induites par toutes les questions binaires qui peuvent être générées à partir des variables descriptives. On utilise donc, ici, l'approche monothétique des arbres de décisions et de régression [2] mais dans un cadre non supervisé. Les différences sont nombreuses et, en particulier, il n'y a pas de variable à expliquer et pas d'élagage.

La méthode DIVCLUS-T [3] est une méthode simple qui suit les principes de la méthode DIV et dont la principale propriété est de fournir une interprétation immédiate et monothétique du dendrogramme et des classes de la hiérarchie.

Dans ce papier, nous proposons une extension de DIVCLUS-T, appelée C-DIVCLUS-T qui est capable de prendre en compte les contraintes de contiguïté. Pour cela nous allons définir un nouveau critère d'homogénéité à partir de la distance mais qui inclus aussi ces contraintes de contiguïté. De ce fait C-DIVCLUS-T sera en mesure de traiter des données complexes. Plusieurs travaux introduisant contraintes de contiguïté dans la méthode de classification automatique ont déja été réalisés [6], [7]. La méthode proposée ici à la spécificité d'être monothétique.

2 Définition du nouveau critère d'homogénéité

Soit $P_K = \{C_1, \ldots, C_k, \ldots, C_K\}$ une partition de K classes de Ω et $\mathbf{D} = (d_{ii'})_{n\times n}$ la matrice des distances. Sans les contraintes de contiguïté le critère d'homogénéité [4] est défini par :

$$W(P_K) = \sum_{k=1}^{K} D(C_k),$$

avec

$$D(C_k) = \sum_{i\in C_k} \sum_{i'\in C_k} \frac{w_i w_{i'}}{2\mu_k} d^2_{ii'}, \tag{1}$$

et $\mu_k = \sum_{i\in C_k} w_i$. Maintenant nous allons introduire les contraintes géographiques sous la forme de contraintes de contiguïté dans ce critère d'homogénéité.

2.1 Modélisation des contraintes géographiques

Dans notre application les objets de Ω que nous voulons classer sont des zones hydrologiques possédant des contraintes géographiques représentant les zones limitrophes. Nous allons modéliser ces contraintes par un graphe $G = (\Omega, E)$ où E est l'ensemble des

arêtes (i, i') entre deux objets de Ω. Ici il y a une arête entre i et i' si la zone i' est limitrophe de i.

Soit $Q = (q_{ii'})_{n\times n}$ la matrice d'incidence de G où

$$\begin{aligned} q_{ii'} &= 1 \text{ si } (i,i') \in E \,(i \text{ est limitrophe de } i') \\ q_{ii'} &= 0 \text{ sinon.} \end{aligned} \tag{2}$$

2.2 Nouvelle distance entre deux objets

Sans les contraintes le critère d'homogénéité $D(C_k)$ de la classe k est égal à :

$$D(C_k) = \sum_{i\in C_k} \frac{w_i}{2\mu_k} D_i(C_k) \text{ avec } D_i(C_k) = \sum_{i'\in C_k} w_{i'} d^2_{ii'} \tag{3}$$

Avec les contraintes ce critère d'homogénéité $D_i(C_k)$ est modifié et s'écrit maintenant comme :

$$D_i(C_k) = \alpha a_i(C_k) + (1-\alpha) b_i(C_k) \tag{4}$$

avec,

$$a_i(C_k) = \sum_{i'\in C_k} w_{i'}(1 - q_{ii'}) d^2_{ii'} \tag{5}$$

$$b_i(C_k) = \sum_{i'\notin C_k} w_{i'} q_{ii'} (1 - d^2_{ii'}), \tag{6}$$

et $\alpha \in [0, 1]$.

Nous pouvons noter qu'en absence de contraintes la matrice d'incidence Q est une $n \times n$ nulle et que $\tilde{D}_i(C_k) = \alpha D_i(C_k)$. Sinon $\tilde{D}_i(C_k)$ est décomposée en deux parties.

La première partie $a_i(C_k)$ mesure la cohérence entre i et sa classe C_k. Cette valeur est petite quand i est proche des autres objets de la classe $C_k (d_{ii'} \approx 0)$ et que ces objets sont ses voisins $(q_{ii'} = 1)$.

La seconde partie $b_i(C_k)$ mesure l'éloignement de i aux autres classes. Cette valeur est petite quand i est distant des objets n'appartement pas à C_k $(d_{ii'} \approx 1)$ et quand ces objets ne sont pas voisins de i $(q_{ii'} = 0)$.

$$W_\alpha(P_K) = \sum_{k=1}^{K} \sum_{i\in C_k} \frac{w_i}{2\mu_k} (\alpha a_i(C_k) + (1-\alpha) b_i(C_k)). \tag{7}$$

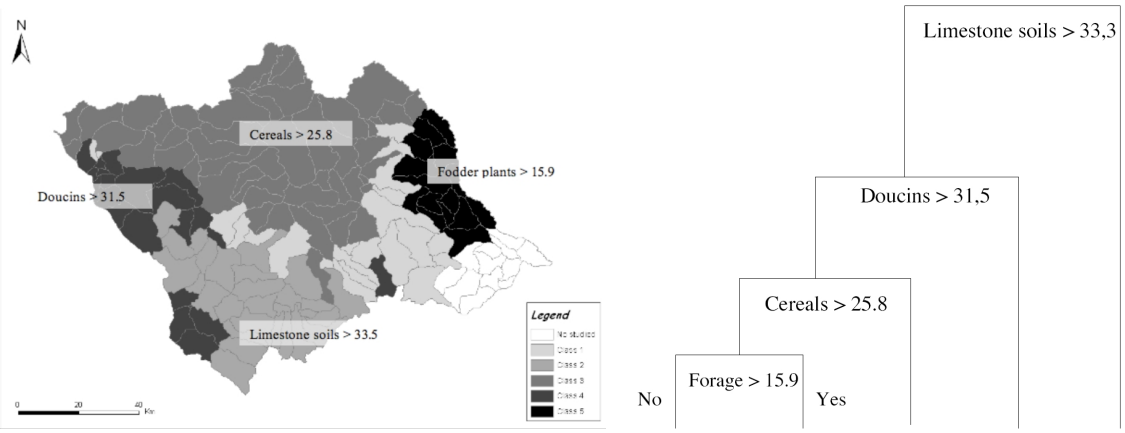

Figure 1: Partition en 5 classes avec $\alpha = 0.5$ associé au dengrogramme obtenu avec C-DIVCLUS-T

3 Application

Les 140 zones hydrologiques sont les unités spatiales et sont connectées entre elles. L'objectif est de partitionner ces zones hydrologiques en classes homogènes par rapport aux variables déscriptives utilisées (variables agricoles ou environnementales) et en essayant de prendre en compte les contraintes géographiques.

Nous avons retenu la partition en 5 classes et la figure 3(a) donne la carte du bassin de la Charente avec cette classification et l'arbre des questions binaires est représenté par la figure3(b).

Nous pouvons observer que les cinq classes sont facilement interprétables. En effet sur la zone côtière les trois classes sont bien délimitées et une région urbaine (deux unités) est mise en valeur.

References

[1] Bock, H.-H. and Diday, E. (eds.): *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data*. Springer Verlag, Heidelberg (2000)

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

[3] Chavent, M, Briant, O. and Lechevallier, Y.: DIVCLUS-T: a monothetic divisive hierarchical clustering method. *Computational Statistics and Data Analysis, 52 (2), 687-701* (2007)

[4] Chavent, M.: Analyse des données symboliques. Une méthode divisive de classification, Thèse de Université Paris-IX Dauphine (1997)

[5] Chavent, M.: A monothetic clustering method. *Pattern Recognition Letters, 19, 989-996* (1998)

[6] Murtagh, F.: A Survey of Algorithm for Contiguity-constrained clustering and Related Problems. *The computer journal, 28(1), 82-88*(1985)

[7] Gordon A. D.: A survey of constrained classification. *Computational statistics and data analysis, 21 (1), 17-29* (1996)

[8] Zahm, F. and Vernier F.: *Contribution to the zoning of territorial agri-environmental measures within the context of the Rural Development Program for the 2007-2013 period: Application of the statistical model RA-SPACE to the river basin district of Adour-Garonne in order to implement a pesticide indicator.* Cemagref report to the French Ministry of Agriculture,122 p (2007)