HAL open science

Monothetic divisive clustering with geographical constraints

Marie Chavent, Yves Lechevallier, Françoise Vernier, Kévin Petit

To cite this version:

Marie Chavent, Yves Lechevallier, Françoise Vernier, Kévin Petit. Monothetic divisive clustering with geographical constraints. Compstat 2008, Aug 2008, Porto, Portugal. pp.67-76. hal-00332024

HAL Id: hal-00332024

https://hal.science/hal-00332024

Submitted on 20 Oct 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Monothetic divisive clustering with geographical constraints ⋆

Marie Chavent[1], Yves Lechevallier[2], Francoise Vernier[3], and Kevin Petit[3]

[1] Université Bordeaux 2, Institut de Mathématiques de Bordeaux, UMR 5251.
146, Rue Léo Saignat, 33076 Bordeaux cedex, France *chavent@sm.u-bordeaux2.fr*
[2] INRIA, Paris-Rocquencourt
78153 Le Chesnay cedex, France *Yves.Lechevallier@inria.fr*
[3] CEMAGREF-Bordeaux, Unité de recherche ADER
50, Avenue de Verdun, 33612 Cestas, France
{*francoise.vernier*}*@bordeaux.cemagref.fr*

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.

Keywords: Divisive clustering, Monothetic cluster, Contiguity constraints

1 Introduction

DIVCLUS-T is a divisive and monothetic hierarchical clustering method which proceeds by optimization of a polythetic criterion (Chavent et al. (2007), Chavent (1998)). The bipartitional algorithm and the choice of the cluster to be split are based on the minimization of the within-cluster inertia. The complete enumeration of all possible bipartitions is avoided by using the same monothetic approach as Breiman et al. (1984) who proposed, and used, binary questions in a recursive partitional process, CART, in the context of discrimination and regression. In the context of clustering, there are no predictors and no response variable. Hence DIVCLUS-T is a DIVisive CLUStering method whose output is not a classification nor a regression tree, but a CLUStering-Tree. Because the dendrogram can be read as a decision tree, it simultaneously provides partitions into homogeneous clusters and a simple interpretation of those clusters.

This algorithm, design for classical data (either categorical or numerical), has also been proposed to deal with more complex data (see Chapter 11.2 of

⋆ Proceedings of the International Conference on Computational Statistics, Compstat'2008

Bock and Diday (2000)). The modification concerns the within-cluster inertia criterion which is replaced by a distance-based criterion and the definitions of binary questions. But with complex data, it is usually not possible to answer directly by *yes* or *no* to a binary question, and the solutions proposed are not always satisfactory.

In this paper we propose an extension of DIVCLUS-T, called C-DIVCLUS-T which is able to take contiguity constraints into account. Because the new criterion defined to include these constraints is a distance-based criterion, C-DIVCLUS-T will be able to deal with complex data. In order to avoid the problem pointed out below concerning the definition of binary questions for complex data, we impose to the variables used in the the binary questions, to be classical. The variables used in the calculation of the distance-based criterion can however have complex descriptions.

Several survey of constrained classification can be found in the literature (see for instance Murtagh (1985), Gordon (1996)). The method proposed here has the specificity to be monothetic and its main advantage is then the simple and natural interpretation of the dendrogram and the clusters of the hierarchy. Of course these monothetic descriptions are also constraints which may deteriorate the quality of the divisions. The price paid by construction in terms of inertia by DIVCLUS-T for this additional interpretation has been studied in Chavent et al. (2007) by applying DIVCLUS-T, Ward and the k-means on six databases from the UCI Machine Learning repository.

In this paper, we present an application of C-DIVCLUS-T to hydrological areas described by agricultural and environmental variables.

2 Definitions and notations

Let $\Omega = \{1, ..., i, ..., n\}$ be of n objects described by p variables $X^1, ..., X^j, ..., X^p$ in a matrix $\mathbf{X}$ of n rows and p columns:

$$\mathbf{X} = (x_i^j) = \begin{array}{c} \\ 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{array}{c} 1 \cdots j \cdots p \\ \begin{bmatrix} \cdot & & \\ \vdots & & \\ \cdots & x_i^j & \cdots \\ \vdots & & \\ \cdot & & \end{bmatrix} \end{array}.$$

For classical data, if X^j is numerical then $x_i^j \in \Re$ and if X^j is categorical then $x_i^j \in M^j$, with M^j the set of categories. For complex data, X^j can be described for instance by an interval $x_i^j = [a_i^j, b_i^j]$ or by a set of categories $x_i^j \subseteq M^j$.

A weight w_i is also associated to each object i. If the data result from random sampling with uniform probabilities, the weights are also uniform :

$w_i = 1$ for all i. It can however be useful for certain applications, to work with non-uniform weights (reweighted sample, aggregate data).

Let V_1 be a subset of $\{X^1, ..., X^j, ..., X^p\}$ with either classical or complex descriptions. Let $\mathbf{D} = (d_{ii'})_{n \times n}$ be a distance matrix with $d_{ii'}$ a distance (or sometimes a dissimilarity) between two objects i and i'. This distances is calculated on the column of $\mathbf{X}$ corresponding to the subset V_1 of variables. In the rest of this paper, we assume that the matrix $\mathbf{D}$ is standardized ($\forall i, i' \in \Omega$, $d_{ii'} \leq 1$) in the following way: If δ is the distance used to compare i and i' on V_1 we have:

$$d_{ii'} = \frac{\delta(i, i')}{\delta_m}, \tag{1}$$

with $\delta_m = \max_{i,i' \in \Omega} \delta(i, i')$. The criterion W, used at each division to evaluate the homogeneity of the bi-partitions, will be defined from $\mathbf{D}$.

Let V_2 be an other subset of $\{X^1, ..., X^j, ..., X^p\}$. As V_1 is used to calculate the matrix distance $\mathbf{D}$ and then the criterion W, the variables in V_2 are used to define at each division, the set of binary questions inducing the finite number of admissible bi-partitions to evaluate. Thanks to the use of binary questions, the computational complexity of the algorithm is reduced and the best bi-partition, chosen according to the criterion W, is monothetic. We recommend to choose in V_2 variables with classical descriptions, such that the binary questions have clear definitions.

We can note that $V_1 \cap V_2$ is not necessarily empty: the same variable can be used to calculate W and the set of binary questions.

3 DIVCLUST-T algorithm

The goal of DIVCLUST-T algorithm is to split recursively a cluster into two sub-clusters, the algorithm starts from the set of objects Ω and the splitting process is stopped after a number of iterations which may be specified by the user. The output of this divisive clustering algorithm is an indexed hierarchy (dendrogram) which is also a decision tree. More precisely at each recursive step, the descendant hierarchical clustering algorithm DIVCLUS-T:

- splits a cluster C_ℓ into a bipartition $(A_\ell, \bar{A}_\ell)$ which minimizes the distance-based criterion W. In Edward and Cavalli-Sforza (1965) method one chooses the optimal bipartition $(A_\ell, \bar{A}_\ell)$ of C_ℓ among all $2^{n_i - 1}$ possible partitions where n_i is the number of objects belonging C_ℓ. It is clear thant the amount of calculation needed when n_i is large will be prohibitive. DIVCLUST approach reduce the complexity by choosing the best bipartition among all the bipartitions induced by a set of all possible binary questions.
- chooses in the partition P_k the cluster C_ℓ to be split in such a way that the new partition P_{k+1} minimizes the distance-based criterion W.

In the complex data context the difficulty is to define (see Bock and Diday (2000))a distance on the set of complex variables included in the set V_1. In the following chapter we propose a new distance-based criterion where the geographical contraints are added to the initial distances without changing their calculation.

The binary questions on a numerical or categorical variables of the set V_2 are easily defined (Chavent et al. (2007)). Some approaches, decribed in Chavent (1998) or in the chapter 11.2 of Bock and Diday (2000), give many strategies to construct a set of binary questions on the complex variables included in the set V_2.

4 A distance-based criterion

Let $P_K = \{C_1, \ldots, C_k, \ldots, C_K\}$ be a K-clusters partition of Ω and $\mathbf{D} = (d_{ii'})_{n\times n}$ the distance matrix. A distance-based homogeneity criterion can be defined as:

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

with

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

and $\mu_k = \sum_{i\in C_k} w_i$.

In case of numerical data with uniform weights compared with the Euclidean distance, $W(P_K)$ is the well-known within-clusters sum of squares criterion.

This distance-based criterion has the advantage to avoid centroids, often difficult to define explicitly in case of complex data. But because of the double sum in its definition, it has the drawback to increase the computational complexity.

Let now introduce geographical constraints in this criterion.

4.1 The geographical constraints

In the real application studied in this paper, the objects of Ω have geometrical constraints. Generally speaking, spatial constraints can be represented in a graph $G = (\Omega, E)$ where E is a set of edges (i, i') between two objects of Ω. There will be an edge between i and i' if i' is a neighbor of i.

Let $Q = (q_{ii'})_{n\times n}$ be the adjacency matrix of G where

$$\begin{aligned} q_{ii'} &= 1 \text{ if } (i, i') \in E \,(i' \text{ is a neighbor of } i') \\ q_{ii'} &= 0 \text{ otherwise.} \end{aligned} \tag{3}$$

4.2 The new distance-based criterion

The criterion $D(C_k)$ can be decomposed in the following way:

$$D(C_k) = \sum_{i \in C_k} \frac{w_i}{2\mu_k} D_i(C_k) \text{ where } D_i(C_k) = \sum_{i' \in C_k} w_{i'} d_{ii'}^2 \tag{4}$$

The criterion $D_i(C_k)$ measures the proximity (dissimilarity) between the object i and the cluster C_k to which it belongs.

In order to take the geographical constraints into account, the criterion $D_i(C_k)$ is modified and re-written in the following way:

$$\tilde{D}_i(C_k) = \alpha a_i(C_k) + (1-\alpha) b_i(C_k) \tag{5}$$

with,

$$a_i(C_k) = \sum_{i' \in C_k} w_{i'} (1 - q_{ii'}) d_{ii'}^2 \tag{6}$$

$$b_i(C_k) = \sum_{i' \notin C_k} w_{i'} q_{ii'} (1 - d_{ii'}^2), \tag{7}$$

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

First we can notice that in the absence of constraints, the adjacency matrix Q is a $n \times n$ null matrix and that $\tilde{D}_i(C_k) = \alpha D_i(C_k)$. Otherwise $\tilde{D}_i(C_k)$ is decomposed into two parts. The first part $a_i(C_k)$ measures the coherence between i and its cluster C_k. It it small when i is similar to the objects in $C_k (d_{ii'} \approx 0)$ and when these objects are neighbor ($q_{ii'} = 0$) of i. The second part $b_i(C_k)$ measures the coherence between i and objects in other clusters than C_k. It is small when i is dissimilar from the objects not in C_k ($d_{ii'} \approx 1$) and when these objects are not neighbors of i ($q_{ii'} = 0$).

In other words, $a_i(C_k)$ measures of a dissimilarity between i and C_k by assigning the value 0 for the neighbors of i and the square of the distance for the other objects belonging to the clusters of i. The second part $b_i(C_k)$ represents a penalty for the neighbors of i which belongs to other clusters.

The new distance-based criterion taking the constraints into account in then:

$$\tilde{W}_\alpha(P_K) = \sum_{k=1}^{K} \sum_{i \in C_k} \frac{w_i}{2\mu_k} (\alpha a_i(C_k) + (1-\alpha) b_i(C_k)). \tag{8}$$

4.3 Study of the parameter α

The parameter α can be chosen by the user (usually, $\alpha = 0.5$) or defined automatically. In this latter case, the idea is to chose α such that $\tilde{W}_\alpha(P_1) = \tilde{W}_\alpha(P_n)$. Indeed, if $\alpha = 1$,

$$\tilde{W}_1(P_K) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{i' \in C_k} \frac{w_i w_{i'}}{2\mu_k} (1 - q_{ii'}) d_{ii'}^2, \tag{9}$$

and $\tilde{W}_1(P_n) = 0$. If $\alpha = 0$,

$$\tilde{W}_0(P_K) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{i' \notin C_k} \frac{w_i w_{i'}}{2\mu_k} q_{ii'} (1 - d_{ii'}^2), \tag{10}$$

and $\tilde{W}_0(P_1) = 0$.
A compromise is then to take α such that $\tilde{W}_\alpha(P_1) = \tilde{W}_\alpha(P_n)$ which gives:

$$\alpha = \frac{A}{A+B}, \tag{11}$$

and

$$\begin{aligned} A &= \sum_{i \in \Omega} \sum_{i' \in \Omega, i \neq i'} q_{ii'} (1 - d_{ii'}^2), \\ B &= \sum_{i \in \Omega} \sum_{i' \in \Omega} (1 - q_{ii'}) d_{ii'}^2. \end{aligned} \tag{12}$$

5 Hydrological areas clustering

Agricultural policies have recently experienced major reformulations and became more and more spatialised. Defining policy priorities requires appropriate tools (indicators, models) with relevant results about ecological and social features of agricultural practices (CEC, 2001[1]). Agri-environmental indicators (AEIs) provide an essential tool for formalizing information from different sources and to address the impact of agricultural production on the environment. These indicators combine information about agricultural activity and environmental conditions (data on climate, soils, slopes, hydrology, etc.). In order to provide helpful results for decision makers, the statistical information on agricultural activity (mainly at the scale of administrative units) has to be transferred to environmentally relevant entities.

An important political issue is currently the implementation of WFD (Water Framework Directive) in European countries. It stresses that an assessment is required to implement efficient measurement programs to preserve or restore the good ecological status of water bodies. The spatial unity (hydrological unit) corresponds to the water body, which is the elementary partition of aquatic environments selected for the water status assessment.

[1] CEC, 2001. Statistical information needed for the indicators to monitor the integration of environmental concerns into the Common Agricultural Policy. Commission of the European Communities. Communication to the Council and the European Parliament, COM 2001, Brussels

A study is carrying out at Cemagref in the context of the SPICOSA[2] project and of the implementation of WFD: the purpose is to define the relevant spatial unit, helpful for the integrated management of the continuum "Pertuis Charentais Sea" and "Charente river basin". We have to define homogeneous areas within the Charente basin to calculate the spatialised AEIs and to implement an hydrological model (SWAT). The questions are: what type of spatial organization can be used to analyze the impact of agriculture on the freshwaters ? Are WFD existing ones (hydrographic units) relevant? Or should new spatial entities be created ?

In this first step, we decide to use the hydrological area (water bodies) as the relevant elementary spatial unit and to analyse all relevant variables at this scale. There are 140 hydrological units within the studied area. The goal is to partition the hydrological units and to obtain some clusters as homogeneous as possible in order to implement AEIs and the SWAT model. Two major types of variables are considered :

- Variables to characterize agricultural activities: because the territorial limits resulting from the environmental zonings established to support the implementation of the WFD are by construction independent of the French administrative geographical area (rgion, canton, commune), we used first the Ra-space method (Zahm and Vernier (2007)) to perform a spatial analysis of agricultural activities at the scale of the hydrological unit defined in the Water Framework Directive.
- Variables to characterize environmental conditions: some other variables are needed to assess the potential risk of agricultural pesticide or nutrients transfer towards surface waters. These data concern structural sensitivity (slope, soil, distance to river,..). We used GIS tools to intersect geographical layers and calculate the values for these variables at the hydrological unit scale.

The 140 hydrological units are then characterized by 14 types of soils (marshland soils, terraces, valleys, artificial area, lakes, different types of groies, clay soils, doucins, limestone soils, clay-limestone soils, and red lands), 17 types of soil occupation (forest, orchards, vineyard...) , 8 main crops, a mean slope and a drainage rate (sum of the length of rivers within the spatial unit/area of the spatial unit). The file provided by the GIS tools includes the calculation of the percentage of area for each variable (see Table 1 below). A second file, provided also by the GIS tools, includes for each hydological area the list of its neighbors.

The DIVCLUS-T method has been applied to the first data file, and C-DIVCLUS-T has been applied to the same data file taking into account the contiguity of the data given in the neighbors file. The five-clusters partition has been retained in both cases.

[2] SPICOSA project web site: www.spicosa.eu

Zhydro	Type of soil	Soil occupation	Crope	Mean slope	Drainage rate
	S_1 S_2 ... S_{14}	O_1 O_2 ... O_{17}	C_1 C_2 ... C_8		
R000	12 22 ... 7.8	9.8 12.6 ... 9.4	12 8.7 ... 32.1	4.44	11.28
⋮	⋮ ⋮ ⋮ ⋮	⋮ ⋮ ⋮ ⋮	⋮ ⋮ ⋮ ⋮	⋮	⋮

Table 1. The first rows of the data file

Figures 1 and 2 give the map of the Charente basin and the clusters obtained with the two clustering methods for the 140 hydrological units . In order to illustrate the interest of using a monothetic approach for clustering, we have also reported on figure 2 the binary questions of the dendrogram obtained with C-DIVCLUS-T (see figure 3).

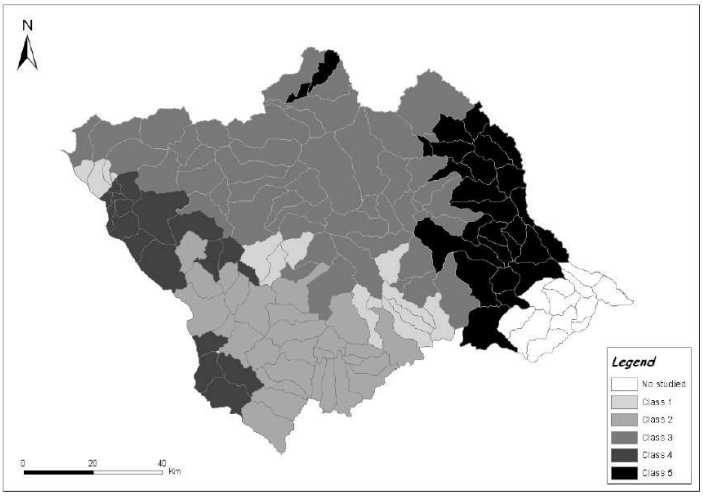

Fig. 1. The five-clusters partition obtained with DIVCLUS-T.

We can observe that the five clusters obtained with C-DIVCLUS-T are more interpretable than those obtained without spatial constraints. Indeed on the coastal zone three clusters are better delimited in figure 2 and a urban area (two hydrological units) is highlighted. Moreover, the hydrological unit of cluster 5 which was alone in the cluster 3 in figure 1, is merged to cluster 3 in figure 2.

Figure 2 can then be read in the following way: a part of the coastal area can be linked to the presence of Doucins soils (moors). In the North of the river basin, an homogeneous area with cereal crops stands out and is not perturbed like in the previous classification. An other relevant area is delimited in the south of the basin with the variable “limestone soils” : we can

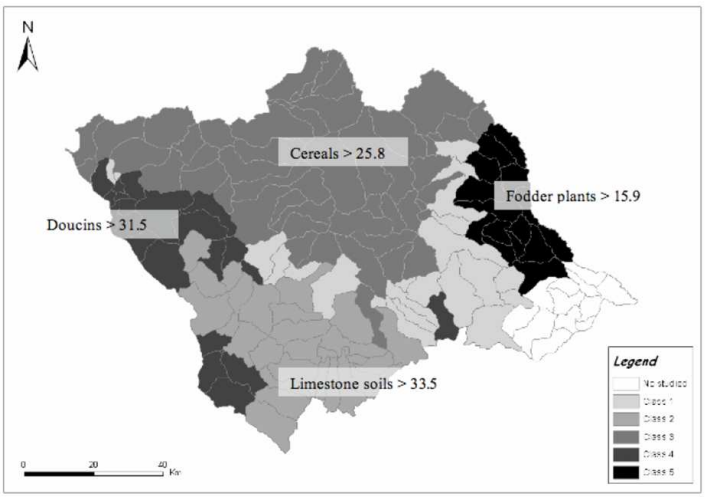

Fig. 2. The five-clusters partition obtained with C-DIVCLUS-T and $\alpha = 0.5$.

find here vineyards and complex cultivation patterns. Finally, the cluster 1 can be linked to more artificialised areas.

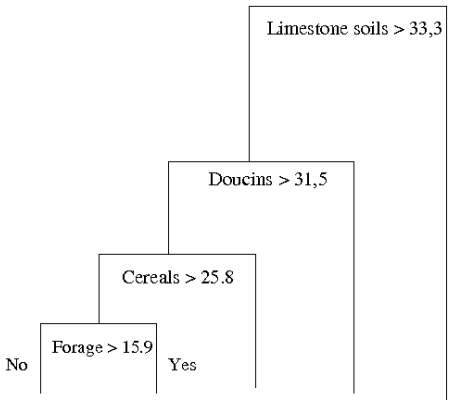

Fig. 3. Dengrogram obtained with C-DIVCLUS-T.

References

BOCK, H.-H. and DIDAY, E. (eds.) (2000): *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data.* Springer Verlag, Heidelberg.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., and STONE, C.J.,(1984): *Classification and regression Trees.* C.A:Wadsworth.

CHAVENT, M, BRIANT, O. and LECHEVALLIER, Y. (2007): DIVCLUS-T: a monothetic divisive hierarchical clustering method. *Computational Statistics and Data Analysis, 52 (2), 687-701.*

CHAVENT, M. (1998): A monothetic clustering method. *Pattern Recognition Letters, 19, 989-996.*

EDWARDS, A.W.F. and CAVALLI-SFORZA, L.L. (1965): A method for cluster analysis. *Biometrics, 21, 362-375.*

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

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

ZAHM, F. and VERNIER F. (2007), *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.