ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

DIVCLUS-T: A monothetic divisive hierarchical clustering method

Marie Chavent[a,*], Yves Lechevallier[b], Olivier Briant[c]

[a] IMB, UMR CNRS 5251, Université Bordeaux1, 351 cours de la libération, 33405 Talence, Cedex, France
[b] INRIA-Rocquencourt, 78153 Le Chesnay, Cedex, France
[c] G-SCOP, ENSGI-INPG, 6 avenue Félix-Viallet, 38031 Grenoble, France

Abstract

DIVCLUS-T is a divisive hierarchical clustering algorithm based on a monothetic bipartitional approach allowing the dendrogram of the hierarchy to be read as a decision tree. It is designed for either numerical or categorical data. Like the Ward agglomerative hierarchical clustering algorithm and the k-means partitioning algorithm, it is based on the minimization of the inertia criterion. However, unlike Ward and k-means, it provides a simple and natural interpretation of the clusters. The price paid by construction in terms of inertia by DIVCLUS-T for this additional interpretation is studied by applying the three algorithms on six databases from the UCI Machine Learning repository.

Keywords: Divisive clustering; Monothetic cluster; Decision dendrogram; Inertia criterion

1. Introduction

The end-point of a classification study is often a partition P of a set of objects Ω into a set of disjoint homogeneous and well-separated clusters $(C_1, \dots, C_k)$. When the desired number of clusters k is "high" the aim of clustering is usually to reduce the number of objects. Each cluster is then replaced by its centroid, and statistical methods can be applied to the centroids weighted by the number of objects in the corresponding cluster. When the number of clusters k is "small" the aim of clustering is usually to find both homogeneous and interpretable clusters. An additional step of cluster interpretation is then necessary.

The idea of this article is to propose a monothetic divisive hierarchical clustering method called DIVCLUS-T. Like the Ward agglomerative hierarchical method and the k-means partitioning method, this divisive method is based on the minimization of the inertia criterion, but it provides, by construction, a simple and natural interpretation of the clusters.

Divisive hierarchical clustering reverses the process of agglomerative hierarchical clustering, by starting with all objects in one cluster, and successively dividing each cluster into smaller ones. A natural approach for dividing a cluster into two non-empty subsets would be to consider all the possible bipartitions. It is clear that such a complete enumeration procedure provides a global optimum but is computationally prohibitive. A variety of divisive clustering

[*] Corresponding author. Tel.: +33 540002116; fax: +33 540002626.
E-mail address: chavent@math.u-bordeaux1.fr (M. Chavent).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.03.013

methods which do not consider all bipartitions have been suggested. MacNaughton-Smith et al. (1964) and Kaufman and Rousseeuw (1990) proposed iterative divisive procedures using an average dissimilarity between an object and a group of objects. Other methods using a dissimilarity matrix as input are based on the optimization of criteria such as the split or the diameter of the bipartition (Guénoche et al., 1991; Wang et al., 1996). For the inertia criterion, divisive counterparts to Ward's agglomerative algorithm have been proposed: for example, instead of splitting by total enumeration it is possible to apply the k-means algorithm, with $k = 2$ (Mirkin, 2005).

In divisive clustering, some methods are polythetic, whereas some others are monothetic. A cluster is called monothetic if a conjunction of logical properties, each one relating to a single variable, is both necessary and sufficient for membership in the cluster (Sneath and Sokal, 1973). A clustering method which builds, by construction, monothetic clusters is then monothetic. In divisive clustering, monothetic clusters are obtained by using, for each division, a single variable and by separating objects having specific variable values from those who do not. Monothetic divisive clustering methods are usually variants of the association analysis method (Williams and Lambert, 1959) and are designed for binary data. We can cite among others Lance and Williams (1968), Kaufman and Rousseeuw (1990). Unlike the first methods cited above, these monothetic methods are not based on the optimization of a "polythetic" criterion like the inertia or the diameter of the bipartitions. These methods are based on the selection, at each stage, of the binary variable which maximizes a measure of association to the other variables. The objects are then divided using the values (0 and 1) of the binary variable.

The divisive clustering method proposed in this paper is monothetic but proceeds by optimization of a polythetic criterion. 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.

In Chavent (1998) a simplified version of DIVCLUS-T was presented for the particular case of quantitative data. Chavent et al. (1999) applied it, together with another monothetic divisive clustering method based on correspondence analysis, to a categorical data set on healthy human skin. A first comparison of DIVCLUS-T with Ward and k-means was given in this paper but only for a single categorical data set and for the 6-cluster partition. More recently, it has been applied to accounting disclosure analysis (Chavent et al., 2005) and a hierarchical divisive monothetic clustering method based on the Poisson process has been proposed in Pircon (2004).

In this paper we present the method DIVCLUS-T in detail (Section 5). This monothetic Ward-like clustering method can also be applied to categorical data and the calculation of the inertia criterion for categorical data is introduced in Section 4. The numerical and categorical examples (Sections 2 and 5.3) show that the main advantage of DIVCLUS-T compared to Ward or k-means is the simple and natural interpretation of the dendrogram and the clusters of the hierarchy. Because these monothetic descriptions are also constraints which may deteriorate the quality of the divisions, in Section 6 we study the price paid by DIVCLUS-T, in terms of inertia, for this additional interpretation. We compare the inertia of the partitions obtained with DIVCLUS-T, Ward and k-means for the six databases of the UCI Machine Learning repository (Hettich et al., 1998).

2. An example

The agglomerative Ward method and the divisive DIVCLUS-T method have been applied to the well-known protein consumption data table (Hand et al., 1994). The dendrogram of the hierarchy obtained with Ward is given in Fig. 1. This dendrogram does not give any information about the interpretation of the clusters such as, for instance, that of {Italy, Greece, Spain, Port}.

The dendrogram of the hierarchy obtained with DIVCLUS-T is given in Fig. 2 and differs from the Ward dendrogram in the inclusion of the monothetic description for each level. For instance, we can read that the countries of the {Italy, Greece, Spain, Port} cluster are characterized by their nuts and fruits/vegetable consumption (Nuts > 3.5) whereas the European countries of the {Fin, Nor, Swed, Den} cluster are characterized by their fish consumption (Fish > 5.7).

The clusters obtained with DIVCLUS-T have natural interpretation but how do the inertias of the clusters obtained by the two methods compare? In the case of the protein data table, all the countries have the same weight and the inertia

Please cite this article as: Chavent, M., et al., DIVCLUS-T: A monothetic divisive hierarchical clustering method. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.03.013

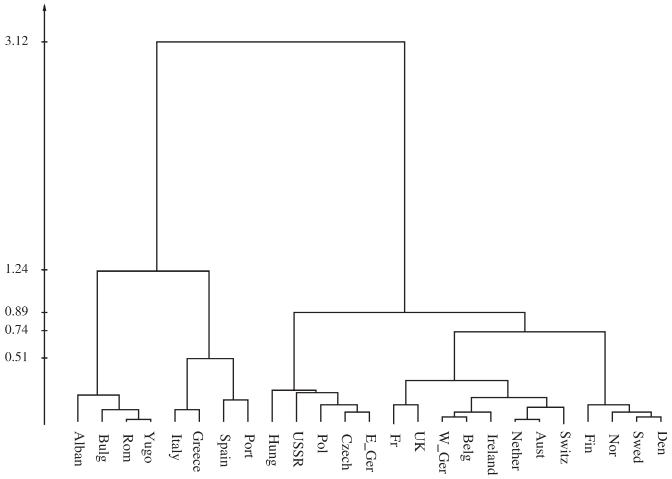

Fig. 1. Ward dendrogram for protein data.

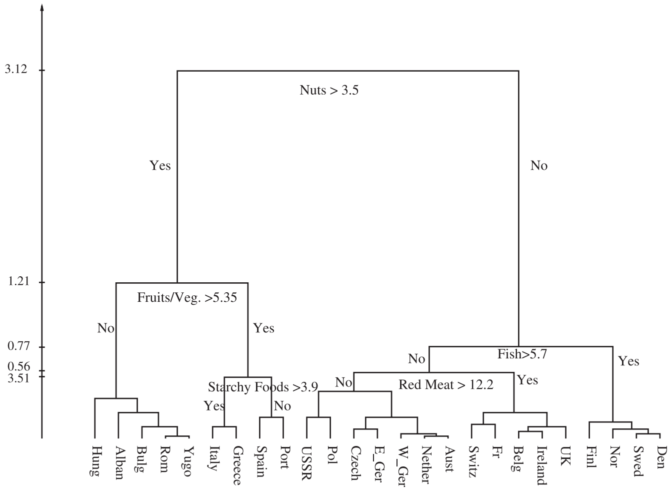

Fig. 2. DIVCLUS-T dendrogram for protein data.

Table 1
Proportion of the inertia (in %) explained by the *k*-cluster partitions obtained with DIVCLUS-T and Ward on the protein data set

k	2	3	4	5	6	7	8	9	10
DIVCLUS-T	37.1	50.6	59.2	65.5	71.2	73.5	79.3	81.6	84
Ward	34.7	48.5	58.5	66.7	72.4	75.5	79	81.6	84

criterion is the classical error sum of squares (SSQ) criterion. We see in Table 1 that the proportion of the explained inertia is better for the partitions of DIVCLUS-T from 2 to 4 clusters and better (or equal) for the partitions of Ward from 5 to 10 clusters. The disadvantage, in term of inertia, of being monothetic seems to be counterbalanced by the advantage of being divisive. Note that few cluster partitions are obtained in the first stages of divisive hierarchical clustering whereas they are obtained in the last stages of agglomerative hierarchical clustering.

3. The data table

We consider a set $\Omega = \{1, \ldots, i, \ldots, n\}$ of n objects which are described by p variables $X^1, \ldots, X^p$ in a matrix $\mathbf{X}$ of n rows and p columns:

$$\mathbf{X} = (x_i^j) = \begin{array}{c} \\ 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{array}{c} 1 \cdots j \cdots p \\ \left[\begin{array}{ccc} & \vdots & \\ & \vdots & \\ \cdots & x_i^j & \cdots \\ & \vdots & \\ & \vdots & \end{array}\right] \end{array}.$$

x_i^j is the value of the jth variable X^j for object i. For a numerical variable, $x_i^j \in \mathbb{R}$ and for a categorical variable, $x_i^j \in M^j$, the set of categories of X^j. We will define by q^j the number of categories of X^j. Here we do not consider the case of a mixed data table and so the $\mathbf{X}$ entries are either all numerical or all categorical.

A weight m_i is associated to each object i and those weights are organized in a vector $\mathbf{m} = (m_1 \cdots m_i \cdots m_n)^{\mathrm{t}}$. If the data result from random sampling with uniform probabilities, the weights are also uniform : $m_i = 1/n$ for all i. But it can be useful, for certain applications, to work with non-uniform weights (reweighted sample, aggregate data).

4. The inertia criterion

A general approach for splitting a set $\Omega = \{1, \ldots, i, \ldots, n\}$ of n objects into k disjoint clusters involves defining a measure of adequacy of a partition P_k and seeking a partition which optimizes that measure. Several possible measures of adequacy exist (Gordon, 1999; Hansen and Jaumard, 1997) and are used in different clustering methods. Here we have chosen to use the inertia criterion (which is a generalization of the error SSQ criterion). Note that a k-clusters partition P_k is a list $(C_1, \ldots, C_k)$ of subsets of Ω verifying $C_\ell \neq \emptyset$ for all $\ell = 1, \ldots, k$, $C_1 \cup \cdots \cup C_k = \Omega$ and $C_\ell \cap C_{\ell'} = \emptyset$ for all $\ell \neq \ell'$.

4.1. Definitions

The inertia criterion is defined on a *numerical* weighted data table $(\mathbf{Z}, \mathbf{w})$ associated with a set Ω of n objects as described in Section 4.2, where

$$\mathbf{Z} = \begin{array}{c} \\ 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{array}{c} 1 \cdots j \cdots p \\ \left[\begin{array}{ccc} & \vdots & \\ & \vdots & \\ \cdots & z_i^j & \cdots \\ & \vdots & \\ & \vdots & \end{array}\right] \end{array}, \quad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_i \\ \vdots \\ w_n \end{bmatrix}.$$

In the calculation of the inertia criterion, an object $i \in \Omega$ will be weighted by w_i and identified with the corresponding row of the matrix $\mathbf{Z}$:

$$\mathbf{z}_i = (z_i^1 \cdots z_i^p)^{\mathrm{t}} \in \mathbb{R}^p.$$

The inertia of a cluster $C_\ell \subseteq \Omega$ is then defined by

$$I(C_\ell) = \sum_{i \in C_\ell} w_i d^2_{\mathbf{M}}(\mathbf{z}_i, \mathbf{g}(C_\ell)), \tag{1}$$

where w_i is the weight of the object i and $\mathbf{g}(C_\ell)$ is the cluster centroid defined by

$$\mathbf{g}(C_\ell) = \frac{1}{\sum_{i \in C_\ell} w_i} \sum_{i \in C_\ell} w_i \mathbf{z}_i.$$

Please cite this article as: Chavent, M., et al., DIVCLUS-T: A monothetic divisive hierarchical clustering method. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.03.013

The distance $d_{\mathbf{M}}$ between the two vectors $\mathbf{z}_i$ and $\mathbf{z}_{i'}$ of $\mathbb{R}^p$ is defined by

$$d^2_{\mathbf{M}}(\mathbf{z}_i, \mathbf{z}_{i'}) = (\mathbf{z}_i - \mathbf{z}_{i'})^{\mathrm{t}}\mathbf{M}(\mathbf{z}_i - \mathbf{z}_{i'}), \tag{2}$$

where $\mathbf{M}$ is a $p \times p$ positive definite matrix.

If $\mathbf{M} = \mathbf{I}$ and $w_i = 1$ for all $i = 1 \cdots n$, then the inertia is the classical error SSQ criterion.

The sum of the inertias of all clusters is called the within-cluster inertia:

$$W(P_k) = \sum_{\ell=1}^{k} I(C_\ell). \tag{3}$$

It is a heterogeneity criterion for the adequacy of a partition $P_k = (C_1, \ldots, C_k)$. Similarly, the inertia of the centroids $\mathbf{g}(C_\ell)$, weighted by $\mu(C_\ell) = \sum_{i \in C_\ell} w_i$, is called the between-cluster inertia:

$$B(P_k) = \sum_{\ell=1}^{k} \mu(C_\ell) d^2_{\mathbf{M}}(\mathbf{g}(C_\ell), \mathbf{g}), \tag{4}$$

with $\mathbf{g} = \mathbf{g}(\Omega)$ the centroid of Ω. This is an isolation criterion for the adequacy of P_k.

Finally, because the total inertia of a set of $\mathbb{R}^p$ points can be partitioned into within- and between-cluster inertia, we have

$$I(\Omega) = W(P_k) + B(P_k), \tag{5}$$

and so minimizing W (the heterogeneity) is equivalent to maximizing B (the isolation).

4.2. The inertia criterion for numerical or categorical data

For a numerical matrix $\mathbf{X}$, the inertia criterion is calculated from the weighted data matrix $(\mathbf{Z}, \mathbf{w})$ with $\mathbf{Z} = \mathbf{X}$ and $\mathbf{w} = \mathbf{m}$. Moreover, the matrix $\mathbf{M}$ used in the quadratic distance $d_{\mathbf{M}}$ defined in (2) is usually the identity matrix $\mathbf{I}$ or the diagonal matrix of the inverse of squared standard deviations:

$$\mathbf{D}_{1/s^2} = \begin{bmatrix} 1/s_1^2 & & 0 \\ & \ddots & \\ 0 & & 1/s_p^2 \end{bmatrix}.$$

This latter distance is used when the variables are measured on very different scales.

For a categorical matrix $\mathbf{X}$, the inertia criterion is calculated from a weighted data matrix $(\mathbf{Z}, \mathbf{w})$ defined as follows. First the matrix $\mathbf{X} = (x_i^j)_{n \times p}$ is converted into an indicator matrix $\mathbf{Q}$ with n rows and q columns, where $q = \sum_{j=1}^{p} q^j$ is the total number of categories in all variables. In each ith row of the indicator matrix, an element is 1 if the object belongs to the corresponding category s of the corresponding categorical variable; otherwise the element is 0. Thus the sum of all the elements in a row is p, the number of variables. A matrix $\mathbf{K} = (k_i^s)_{n \times q}$ is obtained by multiplying for all i the ith row of this indicator matrix $\mathbf{Q}$ by the weight m_i. Because the matrix $\mathbf{K}$ can be considered as a kind of contingency table, the matrix of relative frequencies $\mathbf{F} = (f_i^s)_{n \times q}$ called the correspondence matrix, can be constructed. The relative frequency f_i^s is obtained by dividing the frequency k_i^s by $k_{..} = \sum_i \sum_s k_i^s$, the overall total frequency: $f_i^s = k_i^s / k_{..}$. The second step is to convert the correspondence matrix $\mathbf{F}$ into a row profiles matrix $\tilde{\mathbf{X}}_{n \times q} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n)^{\mathrm{t}}$. The ith row profile $\tilde{\mathbf{x}}_i$ is then defined by dividing the ith row of the correspondence table $\mathbf{F}$ by its marginal total $f_{i.} = \sum_{s=1}^{q} f_i^s$, i.e.

$$\tilde{\mathbf{x}}_i = \left(\frac{f_i^1}{f_{i.}}, \ldots, \frac{f_i^q}{f_{i.}} \right).$$

The total marginals of the correspondence matrix $\mathbf{F}$ are also used to define the row and the column masses: the vector $(f_{1.}, \ldots, f_{i.}, \ldots, f_{n.})$ gives the masses of the rows, and the vector $(f_{.1}, \ldots, f_{.s}, \ldots, f_{.q})$ gives the masses of the columns.

Finally, the inertia criterion is calculated on the weighted data matrix $(\mathbf{Z}, \mathbf{w})$ with $\mathbf{Z}=\tilde{\mathbf{X}}$ and $\mathbf{w}=(f_{1.}, \ldots, f_{i.}, \ldots, f_{n.})^{\mathrm{t}}$. Moreover the matrix $\mathbf{M}$ used in the quadratic distance $d_{\mathbf{M}}$ is not the identity matrix $\mathbf{I}$ or the diagonal matrix $\mathbf{D}_{1/s^2}$ as for numerical data, but the diagonal matrix of the inverse of column masses:

$$\mathbf{D}_{1/f} = diag(1/f_{.1}, \ldots, 1/f_{.q}).$$

The resulting distance

$$d^2_{\mathbf{D}_{1/f}}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_{i'}) = (\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_{i'})^{\mathrm{t}}\mathbf{D}_{1/f}(\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_{i'}),$$

is called the chi-squared distance because, when $\mathbf{K}_{n\times q}$ is a real contingency table crossing a set I of n categories with a set J of q categories, the inertia of the n row profiles weighted by $(f_{1.}, \ldots, f_{i.}, \ldots, f_{n.})$ is identical to chi-square contingency coefficient over $k_{..}$.

5. DIVCLUS-T

In the divisive hierarchical clustering algorithm, one recursively splits a cluster into two sub-clusters, starting from the set of objects $\Omega = \{1, \ldots, n\}$: given the current partition $P_k = (C_1, \ldots, C_k)$, one cluster C_ℓ is split in order to find a partition P_{k+1} which contains $k+1$ clusters and optimizes the chosen adequacy measure, based on the inertia criterion. More precisely, at each stage, the divisive hierarchical clustering method DIVCLUS-T:

- splits a cluster C_ℓ into a bipartition $(A_\ell, \bar{A}_\ell)$ of minimum within-cluster inertia. This bipartitional method is defined in Section 5.1.
- chooses in the partition P_k the cluster C_ℓ to be split in such a way that the new partition P_{k+1} has minimum within-cluster inertia. This choice is explained in Section 5.2, and its link to the construction of the dendrogram is emphasized.

5.1. The problem of how to split a cluster

In order to split optimally a cluster C_ℓ one has to choose the bipartition $(A_\ell, \bar{A}_\ell)$ amongst the $2^{n_\ell - 1} - 1$ possible bipartitions of this cluster of n_ℓ objects. It is clear that such complete enumeration (EdWards and Cavalli-Sforza, 1965) provides a global optimum but is computationally prohibitive. For some adequacy criteria such as the larger of the two sub-clusters diameters, a polynomial-time algorithm exists for the determination of an optimal division (Guénoche et al., 1991).

For the inertia criterion, the k-means iterative relocation algorithms, or one of its several variants (Anderberg, 1973), provide at least one locally optimal division. Here, we have chosen to use a monothetic approach to reduce the number of admissible bipartitions. Breiman et al. (1984) proposed and used binary questions in CART a recursive partitioning process in the context of discrimination and regression. Here we use those binary questions in the context of clustering to reduce the set of possible bipartitions.

5.1.1. Inertia of a bipartition

Let $(A_\ell, \bar{A}_\ell)$ be a bipartition of a cluster C_ℓ of Ω with Ω described by a numerical weighted data table $(\mathbf{Z}, \mathbf{w})$. From (5) we have that minimizing the within-cluster inertia $W(A_\ell, \bar{A}_\ell)$ is equivalent to maximizing the between-cluster inertia $B(A_\ell, \bar{A}_\ell)$. Moreover, we know that $B(A_\ell, \bar{A}_\ell)$ can be written as a weighted distance between the centroids $\mathbf{g}(A_\ell)$ and $\mathbf{g}(\bar{A}_\ell)$:

$$B(A_\ell, \bar{A}_\ell) = \frac{\mu(A_\ell)\mu(\bar{A}_\ell)}{\mu(A_\ell) + \mu(\bar{A}_\ell)} d^2_{\mathbf{M}}(\mathbf{g}(A_\ell), \mathbf{g}(\bar{A}_\ell)). \tag{6}$$

This latter criterion is also the between-cluster distance used in the Ward algorithm.

5.1.2. Binary questions

The binary questions are formulated in terms of the initial data matrix $\mathbf{X}$.

(1) A binary question Q on a numerical variable X^j is given by

Is $X^j \leqslant c$?

This binary question, also denoted by $Q=[X^j \leqslant c]$, splits a cluster C_ℓ into two sub-clusters A_ℓ and $\bar{A}_\ell$ such that $A_\ell=\{i \in C_\ell | x_i^j \leqslant c\}$ and $\bar{A}_\ell=\{i \in C_\ell | x_i^j > c\}$. Because $c \in \mathbb{R}$, the number of binary questions is infinite but these binary questions induce only a finite number of bipartitions $(A_\ell, \bar{A}_\ell)$. Let $x_{(1)}^j, \dots, x_{(i)}^j, \dots, x_{(n_\ell)}^j$ be the ordered values of X^j on the n_ℓ objects of C_ℓ. Obviously the binary questions $[X^j \leqslant c]$ induce the same bipartition $(\{(1), \dots, (i)\}, \{(i+1), \dots, (n_\ell)\})$ for all values of c between two consecutive and different observations $x_{(i)}^j$ and $x_{(i+1)}^j$. By convention and in order to associate a unique binary question to each bipartition $(\{(1), \dots, (i)\}, \{(i+1), \dots, (n_\ell)\})$ the cut values c are defined as the midpoints between two consecutive observations:

$$\left\{ c = \frac{x_{(i)}^j + x_{(i+1)}^j}{2},\ x_{(i)}^j \neq x_{(i+1)}^j,\ \ i = 1, \dots, n-1 \right\}.$$

Thus there will be a maximum of $n_\ell - 1$ different bipartitions induced by the binary questions on X^j.

(2) A binary question Q on a categorical variable X^j is given by

Is $X^j \in M$?

where $M \subset M^j$ is a subset of categories of X^j. This binary question, also denoted by $Q = [X^j \in M]$, splits a cluster C_ℓ into two sub-clusters A_ℓ and $\bar{A}_\ell$ such that $A_\ell = \{i \in C_\ell | x_i^j \in M\}$ and $\bar{A}_\ell = \{i \in C_\ell | x_i^j \in \overline{M}\}$, where $\overline{M}$ is the complement of M in M^j. Let q^j be the number of categories of M^j. If M^j is ordered there are $q^j - 1$ different bipartitions $(M, \overline{M})$ of M^j. Otherwise, there are $2^{q^j-1} - 1$ different bipartitions $(M, \overline{M})$ of M^j. Because the number of bipartitions $(M, \overline{M})$ of M^j is equal to the number of binary questions $[X^j \in M]$, there will be a maximum of $2^{q^j-1} - 1$ different bipartitions $(A_\ell, \bar{A}_\ell)$ of C_ℓ induced by those binary questions. This number of bipartitions grows exponentially with q^j the number of categories.

Up to approximately 13 categories, the totality of bipartitions induced by the variable X^j may be considered for optimization. Beyond that point, a pre-treatment has to be applied to these categories. One possibility (among others) consists of ordering the categories by a preliminary multiple correspondence analysis of the data table $\mathbf{X}$. If p is the number of variables and $q=q^1+\cdots+q^p$ the total number of categories, we consider the coordinates of the q categories in $q - p$ dimensions. Each dimension defines an order for the q categories and also for the q^j categories. There will be $q - p$ different orders of the q^j categories and then at most $(q^j - 1)(q - p)$ different bipartitions $(M, \overline{M})$ of M^j.

Finally, DIVCLUS-T selects the bipartition of maximum between-cluster inertia (defined in (6)) amongst the bipartitions induced by all of the binary questions on all of the p variables $X^1, \dots, X^p$.

5.1.3. *Choice of the binary question*

We now consider the particular case where two binary questions Q and Q' induce two bipartitions $(A_\ell, \bar{A}_\ell)$ and $(A'_\ell, \bar{A}'_\ell)$ of the cluster C_ℓ which maximize the between-cluster inertia. In order to choose amongst these two binary questions we introduce a supplementary criterion D.

(1) For a numerical binary question $Q = [X^j \leqslant c]$ corresponding to the bipartition $(A_\ell, \bar{A}_\ell)$ of C_ℓ this criterion is defined by

$$D(Q) = \frac{B^j(A_\ell, \bar{A}_\ell)}{I^j(C_\ell)}, \tag{7}$$

where the between-cluster inertia $B^j(A_\ell, \bar{A}_\ell)$ and the inertia $I^j(C_\ell)$ are calculated only on the jth column of the matrix $\mathbf{Z} = \mathbf{X}$ (see Section 4.2 for the calculation of the inertia criterion for numerical data). This criterion measures the part of the inertia of the variable X^j explained by the partition $(A_\ell, \bar{A}_\ell)$ of C_ℓ.

(2) For a categorical binary question $Q = [X^j \in M]$ this criterion is defined by

$$D(Q) = \sum_{s \in M^j} \frac{B^s(A_\ell, \bar{A}_\ell)}{I^s(C_\ell)}, \tag{8}$$

Please cite this article as: Chavent, M., et al., DIVCLUS-T: A monothetic divisive hierarchical clustering method. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.03.013

where the between-cluster inertia $B^s(A_\ell, \bar{A}_\ell)$ and the inertia $I^s(C_\ell)$ are calculated on the sth column of the row profile matrix $\mathbf{Z}=\tilde{\mathbf{X}}$ (see Section 4.2 for the calculation of the inertia criterion for categorical data). For the sake of simplicity we write $s \in M^j$ to characterize the column of $\tilde{\mathbf{X}}$ corresponding to categories of X^j. Hence $D(Q)$ is the sum, for all categories of X^j, of the part of the inertia of the category explained by the bipartition $(A_\ell, \bar{A}_\ell)$. $D(Q)$ is the relative contribution of the variable X^j to the bipartition $(A_\ell, \bar{A}_\ell)$.

In both cases, $D(Q)$ measures a discrimination power for the variable X^j with respect to the partition $(A_\ell, \bar{A}_\ell)$. Finally, when two binary questions Q and Q' induce two bipartitions $(A_\ell, \bar{A}_\ell)$ and $(A'_\ell, \bar{A}'_\ell)$ which both maximize the between-cluster inertia, DIVCLUS-T selects the most discriminating one according to this criterion D.

5.2. *Selecting the cluster to be split*

A hierarchy H of Ω is a set of clusters satisfying the following conditions (Gordon, 1999):

(a) $\Omega \in H$;
(b) $\emptyset \notin H$;
(c) the singleton $\{i\} \in H$ for all $i \in \Omega$;
(d) if $A, B \in H$, then $A \cap B \in \{\emptyset, A, B\}$.

An indexed hierarchy is a couple (H, h), where h is a mapping from H to $\mathbb{R}^+$ satisfying the following conditions:

(i) $\forall A \in H$ such that $h(A)=0$, A is a singleton,
(ii) $\forall A, B \in H$, if $A \subset B$, then $h(A) \leqslant h(B)$.

The common graphical representations of indexed hierarchies are dendrograms.

In divisive clustering, the set of clusters obtained after $K-1$ divisions is a hierarchy H_K whose singletons are the K clusters of the partition P_K obtained in the last stage of DIVCLUS-T. In DIVCLUS-T the number $2 \leqslant K \leqslant n$ is given as input by the user. Because the resulting hierarchy can be considered as a partial hierarchy halfway between the top and bottom levels, it is referred to as an upper hierarchy (Mirkin, 2005). This upper hierarchy is indexed by h so that in the dendrogram the height of a cluster C_ℓ split into two sub-clusters A_ℓ and $\bar{A}_\ell$ is (as in Wards method):

$$h(C_\ell) = B(A_\ell, \bar{A}_\ell) = \frac{\mu(A_\ell)\mu(\bar{A}_\ell)}{\mu(A_\ell)+\mu(\bar{A}_\ell)} d^2(g(A_\ell), g(\bar{A}_\ell)).$$

When the divisions are continued until giving singleton clusters (or clusters of objects with identical descriptions), all of the clusters can be systematically split and the full hierarchy H_n can be indexed by h. When the divisions are not continued down to H_n, the clusters are not systematically split: in order to have the dendrogram of the upper hierarchy H_K built by the "top" (the $K-1$ largest) levels of the dendrogram of H_n, a cluster represented higher in the dendrogram of H_n has to be split before the others. DIVCLUS-T then chooses to split the cluster C_ℓ with the maximum value $h(C_\ell)$. Consequently because

$$W(P_{k+1}) = W(P_k) - h(C_\ell),$$

maximizing $h(C_\ell)$ ensures that the new partition $P_{k+1} = P_k \cup \{A_\ell, \bar{A}_\ell\} - \{C_\ell\}$ has a minimum within-cluster inertia. DIVCLUS-T then uses the same idea as Wards agglomerative clustering method.

5.3. *An example for categorical data*

DIVCLUS-T has been applied to a categorical data set where 27 breeds of dogs are described by seven categorical variables (Saporta, 1990). The number of clusters of the finest partition is fixed to $K=8$. The dendrogram of the hierarchy H_8 obtained after seven divisions and the associated seven binary questions are given in Fig. 3.

At the first stage, the divisive clustering method constructs a bipartition of the 27 dogs. There are 13 different binary questions and 13 bipartitions to evaluate: two variables are binary (inducing two bipartitions), four variables are ordinal with three levels (inducing 4×2 bipartitions) and one is nominal with 3 categories (inducing 3 bipartitions). The

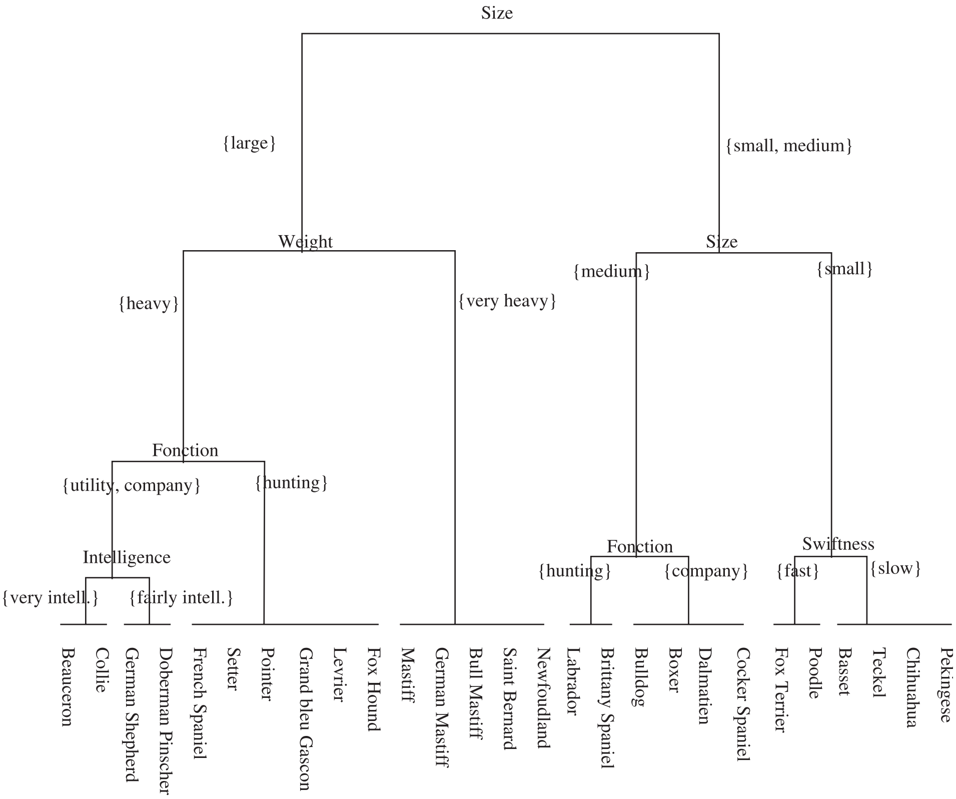

Fig. 3. DIVCLUS-T dendrogram for dogs data.

question “Is size large?” gives the bipartition of smallest within-cluster inertia and is chosen for the first split. For each sub-cluster the “best” bipartition is then obtained in the same way. The between-cluster inertia obtained by splitting the 15 “large” dogs is slightly smaller than that obtained by splitting the 12 “small or medium” dogs and so this latter cluster is divided. This process is repeated until we obtain the final 8-cluster partition.

5.4. *Computational complexity*

For numerical data the computational complexity of DIVCLUS-T is $o(Kpn(\log(n) + p))$, with K the number of clusters of the finest partition, p the number of variables and n the number of objects. Let us briefly compare this complexity with those of Ward and k-means methods. The use of the nearest neighbor method (Mac Quitty, 1966) in agglomerative hierarchical clustering algorithms yields an $o(n^2)$ implementation of the Ward algorithm, the single-linkage, complete linkage and average linkage algorithms (Benzecri, 1982; Murtagh, 1983; Hansen and Jaumard, 1997). This implementation uses a distance matrix as input. When the input is not a distance matrix but a data matrix (objects × variables), the time spent to compute the distance matrix has to be taken into account. For Ward applied to a quantitative data set, because the time spent for the calculation of the $n(n-1)/2$ Euclidean distances is $o(pn^2)$, the complexity is $o(pn^2)$. DIVCLUS-T is then more efficient than Ward for small values of K and $p < n$. That is explained by the fact that divisive algorithms such as DIVCLUS-T need $K-1$ iterations to find the partitions from 2 to K clusters whereas agglomerative algorithms such as Ward need $K-n$ iterations to find the partitions from n to K clusters. The computational complexity of the partitioning k-means algorithm is $o(KpnT)$, where T is the number of iterations (Duda et al., 2001). DIVCLUS-T is also more efficient than k-means when $\log(n) + p < T$.

For categorical data, finding the best bipartition induced by the binary questions can be computationally expensive. In DIVCLUS-T we propose a preliminary treatment which consists of ordering the categories by multiple correspondence analysis of the data table **X** (see Section 5.1.2). By taking this order into account, the number of binary questions on a qualitative variable X^j with q^j categories decreases from $2^{q^j-1}-1$ to q^j-1 and consequently the set of bipartitions evaluated at each stage is also reduced. The combinatorial problem is then reduced but note that the quality of the best bipartition may be degraded in terms of inertia if the set of possible bipartitions is too small. For qualitative variables with less than approximately 13 categories, complete enumeration is preferred. For ordinal variables there is no combinatorial problem because of the natural order of the categories, but the problem of the quality of the best bipartition in a small set of possible bipartition remains.

6. Empirical comparison with Ward and *k*-means

The advantage of DIVCLUS-T in comparison to Ward or k-means clustering methods is the direct and natural interpretation of the clusters. However, what is the price paid by DIVCLUS-T, in terms of inertia, for this additional monothetic description of the clusters? In order to answer this question, we have applied DIVCLUS-T, Ward and k-means algorithms to three numerical and three categorical data sets of the UCI Machine Learning repository (Hettich et al., 1998). A short description of the six data sets is given in Table 2.

6.1. The proportion of inertia explained

The method DIVCLUS-T uses:

- the matrix $\mathbf{X}$ to calculate the binary questions and then the set of possible bipartitions at each stage,
- a weighted data matrix $(\mathbf{Z}, \mathbf{w})$ and a distance $d_{\mathbf{M}}$ to calculate the inertia criterion of those bipartitions (see Section 4.2):
    - for the three numerical data sets, $\mathbf{Z}=\mathbf{X}$, $\mathbf{w}=(1,\ldots,1)^{\mathrm{t}}$ and $\mathbf{M}=\mathbf{D}_{1/s^2}$ the diagonal $p \times p$ matrix of the inverse of squared standard deviations,
    - for the three categorical data sets, $\mathbf{Z}=\tilde{\mathbf{X}}$ the row profiles matrix, $\mathbf{w}=(1/n,\ldots,1/n)^{\mathrm{t}}$ (because $m_i=1$ and then $f_{i.}=1/n$) and $\mathbf{M}=\mathbf{D}_{1/f_{.s}}$ the diagonal $q \times q$ matrix of the inverse of column masses.

The Ward and k-means clustering methods use the same weighted data matrix $(\mathbf{Z}, \mathbf{w})$ and distance $d_{\mathbf{M}}$ than DIVCLUS-T, but do not use the raw data matrix $\mathbf{X}$ within the clustering algorithm.

The three clustering methods all use the inertia criterion. Hence the quality of the partitions P_k built by the three methods from the same set of objects Ω, described by the weighted data matrix $(\mathbf{Z}, \mathbf{w})$, can be ranked using the proportion E of inertia explained by

$$E(P_k) = 100 \times \left(1 - \frac{W(P_k)}{I(\Omega)}\right). \tag{9}$$

This lies between 0% and 100% and is equal to 100% for the singleton partition and to 0% for the single cluster (Ω) partition. Because E increases with the number of clusters k of the partition, it can be used only to compare partitions having the same number of clusters. In the following, we assume that a partition P_k is better than a partition P'_k if $E(P_k) > E(P'_k)$.

Table 2
Data set descriptions

Name	Type	# Objects	# Variables (# categories)
Glass	Numerical	214	8
Pima Indians diabetes	Numerical	768	8
Abalone	Numerical	4177	7
Zoo	Categorical	101	$15(2)+1(6)$
Solar flare	Categorical	323	$2(6)+1(4)+1(3)+6(2)$
Contraceptive method choice (CMC)	Categorical	1473	$9(4)$

Please cite this article as: Chavent, M., et al., DIVCLUS-T: A monothetic divisive hierarchical clustering method. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.03.013

Table 3
Numerical data sets: proportion $E(P_k)$ of inertia explained

k	Glass			Pima			Abalone		
	DIV	Ward	k-means	DIV	Ward	k-means	DIV	Ward	k-means
2	21.5	22.5	22.8	14.8	13.3	16.5	60.2	57.7	60.9
3	33.6	34.1	35.0	23.2	21.6	29.0	72.6	74.8	76.0
4	45.2	44.3	46.6	29.4	29.4	36.2	81.8	80.0	82.6
5	53.4	53.0	54.7	34.6	34.9	41.0	84.2	85.0	86.1
6	58.2	58.4	60.7	38.2	40.0	45.3	86.3	86.8	87.9
7	63.1	63.5	65.7	40.9	44.4	48.9	88.3	88.4	89.6
8	66.3	66.8	68.2	43.2	47.0	51.2	89.8	89.9	90.9
9	69.2	69.2	70.5	45.2	49.1	53.2	91.0	90.9	91.8
10	71.4	71.5	72.4	47.2	50.7	55.1	91.7	91.6	92.4
11	73.2	73.8	74.7	48.8	52.4	56.7	92.0	92.1	92.8
12	74.7	75.9	76.6	50.4	53.9	58.4	92.3	92.4	93.1
13	76.2	77.6	77.2	52.0	55.2	59.7	92.6	92.7	93.4
14	77.4	79.1	78.2	53.4	56.5	61.1	92.8	93.0	93.7
15	78.5	80.4	79.3	54.6	57.7	62.1	93.0	93.2	93.9

Table 4
Categorical data sets: proportion $E(P_k)$ of inertia explained

k	Zoo			Solar Flare			CMC		
	DIV	Ward	k-means	DIV	Ward	k-means	DIV	Ward	k-means
2	23.7	22.4	23.7	12.7	12.6	12.7	8.4	7.6	9.1
3	38.2	37.1	38.2	23.8	22.4	23.8	14.0	12.8	15.3
4	50.1	50.3	51.1	32.8	29.3	33.1	18.9	17.3	19.9
5	55.6	55.8	56.5	38.2	35.1	38.6	23.0	21.5	23.9
6	60.9	61.1	61.0	43.0	40.1	43.2	26.3	25.2	27.7
7	65.6	65.5	66.1	47.7	45.0	48.0	28.4	28.5	30.9
8	68.9	68.6	67.1	51.6	49.8	52.1	30.3	30.9	33.7
9	71.8	71.5	69.7	54.3	53.5	55.8	32.1	33.2	36.2
10	74.7	74.4	73.1	57.0	57.1	58.6	33.8	35.5	38.5
11	76.7	76.6	72.9	59.3	60.4	61.4	35.5	37.4	40.3
12	78.4	78.6	77.7	61.3	62.9	64.1	36.9	39.1	42.3
13	80.1	80.1	77.8	63.1	65.2	65.6	38.1	40.6	43.2
14	81.5	81.4	80.1	64.5	67.2	66.6	39.2	42.1	44.4
15	82.7	82.6	81.0	65.8	68.6	68.6	40.3	43.5	46.1

Tables 3 and 4 give partitions from 2 to 15 clusters for the numerical and categorical data sets, respectively. For each data set the two first columns display the proportion of inertia explained for partitions built with DIVCLUS-T and Ward. The third column displays the proportion of inertia explained for the *k*-means partitions.

First we compare the results for the three numerical data sets (see Table 3). For the Glass data set the partitions obtained with DIVCLUS-T are either better (for 4 clusters), worse (for 2, 3, and from 12 to 15 clusters) or equivalent (from 5 to 11 clusters). For the Pima data set the partitions obtained with DIVCLUS-T are better or equivalent up to 4 clusters, and Ward takes the lead from 5 clusters onwards. Because DIVCLUS-T is divisive whereas Ward is agglomerative, it is not really surprising that Ward tends to become better than DIVCLUS-T as the number of clusters increases. For the Abalone data set which is bigger than the two previous ones (4177 objects), DIVCLUS-T behaves better than Ward until 3 clusters and the results are very close afterwards. One reason for having better results with DIVCLUS-T on the Abalone data set is probably the larger number of objects in this database. Indeed the number of bipartitions considered for optimization at each stage increases with the number of objects. We can then expect better results with larger data sets.

Please cite this article as: Chavent, M., et al., DIVCLUS-T: A monothetic divisive hierarchical clustering method. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.03.013

Table 5
Subsample descriptions

Data set	Data set size n	Proportion α	Subsample size N
Glass	214	0.70	150
Pima	768	0.65	500
Abalone	4177	0.12	500
Zoo	101	0.79	80
Solar	323	0.62	200
CMC	1473	0.27	400

Table 6
Percentage of subsamples where DIVCLUS-T performs better than Ward

k	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Glass	36	56	59	57	48	41	42	34	25	11	7	5	3	2
Pima	91	36	15	5	1	0	0	0	0	0	0	0	0	0
Abalone	71	21	83	37	22	28	65	50	25	12	7	3	1	1
Zoo	96	88	47	39	43	43	48	63	64	48	47	44	46	43
Solar	91	100	99	100	98	92	71	44	15	2	0	0	0	0
CMC	98	99	100	98	89	61	40	19	9	3	4	1	1	1

For the third column, the k-means algorithm is executed 100 times with different initial seeds and the best solution is retained. Because this algorithm locally optimizes the within-cluster inertia, it is to be expected that the proportion of inertia explained will generally be greater than for the other methods. Finally, for these three continuous data sets, DIVCLUS-T seems to perform better for few cluster partitions and for larger data sets.

For the three categorical data sets (see Table 4) we obtain the same kind of results. For the Solar Flare and CMC data sets DIVCLUS-T is better than Ward until 10 and 8 clusters, respectively. For the Zoo data set, DIVCLUS-T performs worse than Ward; this may be because all the variables in the Zoo data set are binary and, as stated before, the quality of the results (in terms of inertia) depends upon the number of categories and variables.

These worse results (in terms of inertia) with DIVCLUS-T were to be expected because of the constraint for the clusters to be monothetic. We can, however, conclude from these examples that in spite of this constraint, DIVCLUS-T performs quite well for few cluster partitions, probably because these arise in the first steps of DIVCLUS-T, whereas they come from the last steps of Ward.

6.2. *Resampling*

We have used a resampling procedure to go further in the comparison of the three clustering methods. Details on resampling procedures for validation can be found in Mirkin (2005). In our application, the resampling procedure includes the following steps:

A. Generation of a number of data sets, *copies*, by subsampling: a proportion α, $0 < \alpha < 1$, is specified and αn objects are randomly selected without replacement as a subsample; data consists of rows corresponding to selected objects. Table 5 gives for each data set, the size n, the proportion α and the subsample size $N = \alpha n$. Hundred subsamples have been generated for each data set.
B. Running successively the three clustering algorithms for all 600 copies.
C. Evaluating the results for each individual copy: For this, two indices are determined for the partitions from 2 to 15 clusters:
   - the first index is equal to 1 if the proportion of inertia explained for the partition obtained with DIVCLUS-T is greater than that obtained with Ward. It is equal to 0 otherwise.

Please cite this article as: Chavent, M., et al., DIVCLUS-T: A monothetic divisive hierarchical clustering method. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.03.013

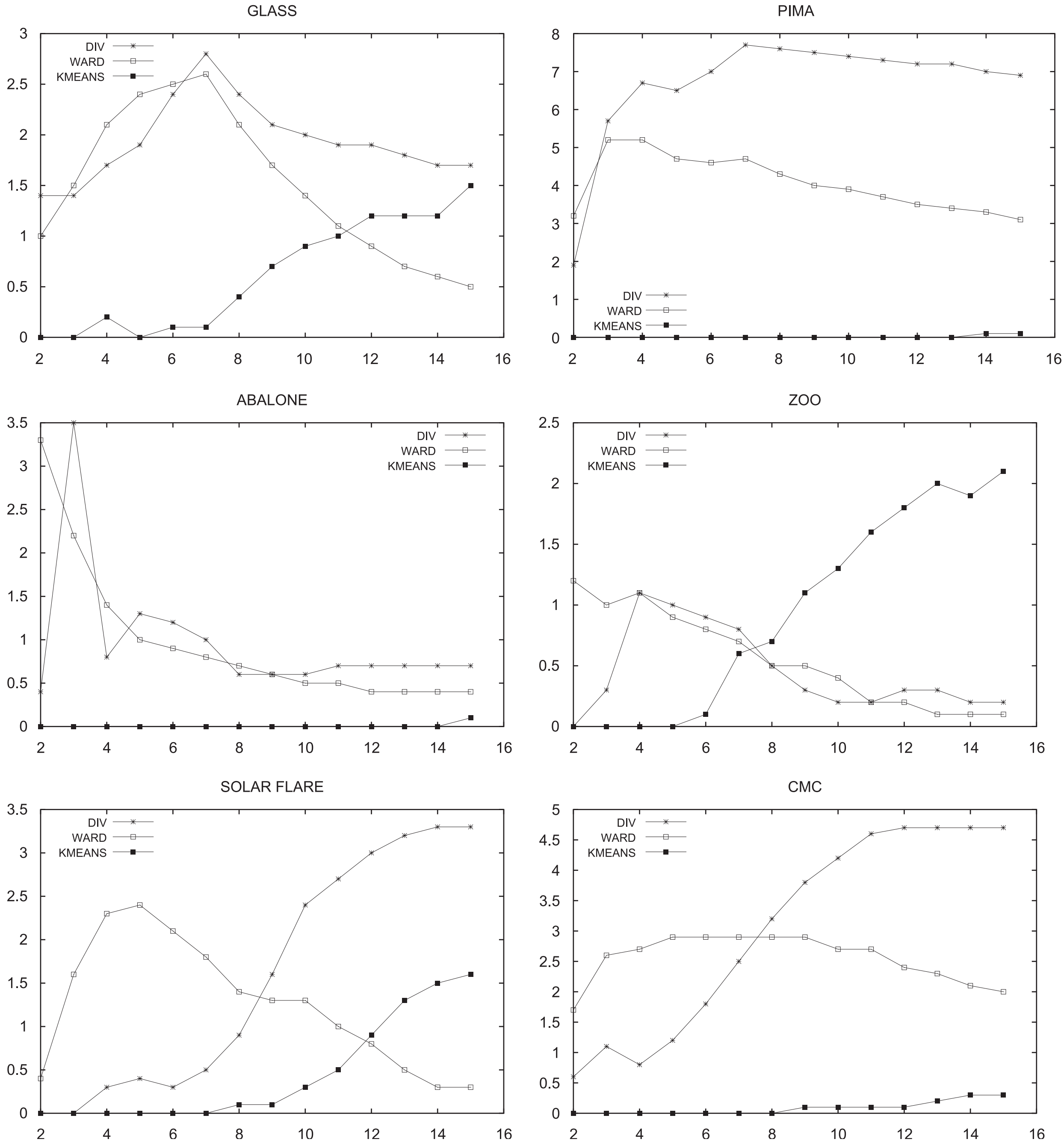

Fig. 4. Deviations from the best solution as proportion of explained inertia.

- the second index is, for each clustering method, the difference between the largest of the three proportions of inertia explained for partitions obtained with the three methods, and the proportion of inertia explained for the partition obtained with the particular method.

D. Aggregating results: For each data set and for partitions from 2 to 15 clusters, the 100 evaluations for each individual copy are averaged. The proportion of subsamples where DIVCLUS-T is better than Ward (in terms of inertia) is calculated by averaging the first index values, see Table 6. The mean deviation of each clustering method from the best solution is given by averaging the second index values, see Fig. 4.

Please cite this article as: Chavent, M., et al., DIVCLUS-T: A monothetic divisive hierarchical clustering method. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.03.013

Table 6 confirms that DIVCLUS-T performs relatively well, despite the monothetic constraint, for few cluster partitions. For the CMC data set for instance, the two-cluster partitions of DIVCLUS-T is better than two-cluster partitions of Ward on 98 subsamples out of 100.

Fig. 4 shows that the deviation from the best solution is, in the worst case, equal to 8% of explained inertia (for the PIMA data set, seven clusters and DIVCLUS-T). Because the k-means algorithm locally optimizes the inertia criterion, most of the time it gives the best solution. If the k-means partition is the best of the three partitions for the 100 subsamples, its mean deviation is equal to 0. This mean deviation is always close to 0 for the Abalone, Pima and CMC data sets. It is noticeable, however, that for the other three data sets (Glass, Zoo, Solar Flame) the k-means solution is not always the best and that its deviation from the best solution increases with the number of clusters. Ward even takes the lead from a certain number of clusters onwards. DIVCLUS-T and Ward curves evolve the same way for Glass, Pima, Abalone and Zoo data sets, with Ward above DIVCLUS-T when the number of clusters increases. The deviations of Ward from the best solution are not so different from the deviations of DIVCLUS-T from the best solution for the Abalone and Zoo data sets (except for few cluster partitions). To sum up, it is difficult to conclude from these results that one clustering method is much better than the others in terms of inertia. DIVCLUS-T is not systematically clearly worse than the two other clustering methods. This is, in itself, a reassuring result for this clustering method which allows, by construction, a simple interpretation of the clusters.

7. Conclusion

This paper proposes a divisive monothetic hierarchical clustering method designed for either numerical or categorical data. For categorical data, the inertia criterion is calculated by converting the data table into a numerical correspondence matrix and a row-profiles matrix. The χ^2 distance is then used in place of Euclidean distance. A solution to the computational problem for categorical binary questions is also proposed. The advantage of DIVCLUS-T, compared to classical methods like Ward or k-means, is the direct interpretation of the clusters: the hierarchy can be read as a decision tree. Of course this advantage has to be balanced with a relative rigidity of the clustering process. Some specific simulations should be able to show easily that DIVCLUS-T is unable to find correctly clusters of specific shapes. But what are the shapes of the clusters in real data sets? We have also seen, on the six data sets from the UCI Machine Learning repository, that the price paid, in terms of inertia, by DIVCLUS-T for the interpretability of the clusters is not systematic, especially for few cluster partitions. Finally, if the user is interested in rather large partitions, for example in order to reduce the number of objects, Ward and k-means are certainly more efficient than DIVCLUS-T. However, if the user is interested in few cluster partitions with good interpretation, DIVCLUS-T seems to be an interesting alternative to classical methods.

References

Anderberg, M.R., 1973. Cluster Analysis for Applications. Academic Press, New York.

Benzecri, J.P., 1982. Construction d'une classification ascendante hiérarchique par la recherche de chaine des voisins réciproques. Les Cahiers de l'Analyse des Données VII (2), 209–219.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

Chavent, M., 1998. A monothetic clustering method. Pattern Recognition Lett. 19, 989–996.

Chavent, M., Guinot, C., Lechevallier, Y., Tenenhaus, M., 1999. Méthodes divisives de classification et segmentation non supervisée: recherche d'une typologie de la peau humaine saine. Rev. Statist. Appl. XLVII (4), 87–99.

Chavent, M., Ding, Y., Fu, L., Stolowy, H., Wang, H., 2005. Disclosure and determinants studies: an extension using the divisive clustering method (DIV). European Accounting Rev. 15 (2), 181–218.

Duda, R.O., Hart, P.E., Strok, D.G., 2001. Pattern Classification Wiley-Interscience, second ed.

EdWards, A.W.F., Cavalli-Sforza, L.L., 1965. A method for cluster analysis. Biometrics 21, 362–375.

Gordon, A.D., 1999. Classification. second ed. Chapman & Hall, CRC Press, London, Boca Raton, FL.

Guénoche, A., Hansen, P., Jaumard, B., 1991. Efficient algorithms for divisive hierarchical clustering with the diameter criterion. J. Classification 8, 5–30.

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J., Ostrowski, E. (Eds.), 1994. A Handbook of Small Data Sets. Chapman & Hall, London.

Hansen, P., Jaumard, B., 1997. Cluster analysis and mathematical programming. Math. Programming 4, 215–226.

Hettich, S., Blake, C.L., Merz, C.J., 1998. UCI repository of machine learning databases. ⟨http://www.ics.uci.edu/mlearn/MLRepository.html⟩, University of California, Department of Information and Computer Science, Irvine, CA.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data. Wiley, New York.

Lance, G.N., Williams, W.T., 1968. Note on a new information statistic classification program. Comput. J. 11, 195–197.

MacNaughton-Smith, P., Williams, W.T., Mockett, L.G., 1964. Dissimilarity analysis: a new technique of hierarchical subdivision. Nature 202, 1034–1035.

Mac Quitty, L.L., 1966. Similarity analysis by reciprocal pairs of discrete and continuous data. Ed. Psych. Meas. 26, 825–831.

Mirkin, B., 2005. Clustering for Data Mining. A Data Recovery Approach. Chapman & Hall, CRC Press, London, Boca Raton, FL.

Murtagh, F., 1983. A survey of recent advances in hierarchical clustering algorithms. Comput. J. 26, 329–340.

Pircon, J.-Y., 2004. La classification et les processus de Poisson pour de nouvelles méthodes de partitionnement. Ph.D. Thesis, Facultés Universitaires Notre-Dame de la Paix, Namur, Belgium.

Saporta, G., 1990. Probabilités Analyse des données et Statistique. Editions TECHNIP.

Sneath, P.H., Sokal, R.R., 1973. Numerical Taxonomy. Freeman, San Francisco.

Wang, Y., Yan, H., Sriskandarajah, C., 1996. The weighted sum of split and diameter clustering. J. Classification 13, 231–248.

Williams, W.T., Lambert, J.M., 1959. Multivariate methods in plant ecology. J. Ecology 47, 83–101.

Please cite this article as: Chavent, M., et al., DIVCLUS-T: A monothetic divisive hierarchical clustering method. Comput. Statist. Data Anal. (2007), doi: 10.1016/j.csda.2007.03.013