HAL
open science

Divisive Monothetic Clustering for Interval and Histogram-valued Data

Paula M. Brito, Marie Chavent

To cite this version:

Paula M. Brito, Marie Chavent. Divisive Monothetic Clustering for Interval and Histogram-valued Data. ICPRAM 2012 - 1st International Conference on Pattern Recognition Applications and Methods, Feb 2012, Portugal. pp.229-234. hal-00763215

HAL Id: hal-00763215

https://hal.science/hal-00763215v1

Submitted on 10 Dec 2012

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.

DIVISIVE MONOTHETIC CLUSTERING FOR INTERVAL AND HISTOGRAM-VALUED DATA

Paula Brito[1] and Marie Chavent[2]

[1]*Faculdade de Economia & LIAAD-INESC Porto LA, Universidade do Porto, Portugal*

[2]*UFR Sciences et Modélisation, IMB et INRIA CQFD, Université de Bordeaux 2, Bordeaux, France*

mpbrito@fep.up.pt, Marie.Chavent@math.u-bordeaux1.fr

Keywords: Divisive clustering, Histogram data, Interval data, Monothetic clustering.

Abstract: In this paper we propose a divisive top-down clustering method designed for interval and histogram-valued data. The method provides a hierarchy on a set of objects together with a monothetic characterization of each formed cluster. At each step, a cluster is split so as to minimize intra-cluster dispersion, which is measured using a distance suitable for the considered variable types. The criterion is minimized across the bipartitions induced by a set of binary questions. Since interval-valued variables may be considered a special case of histogram-valued variables, the method applies to data described by either kind of variables, or by variables of both types. An example illustrates the proposed approach.

1 INTRODUCTION

Clustering is a multivariate data analysis technique aiming at organizing a set of entities in a family of clusters on the basis of the values observed on a set of descriptive variables. Hierarchical clustering provides a set of nested partitions, ranging from the trivial one with one element per cluster to that consisting of one single cluster gathering all entities together. Agglomerative algorithms proceed bottom-up, merging at each step the two most similar clusters until the cluster containing all entities is formed, while divisive algorithms proceed top-down, starting with all entities in one single cluster, and perform a bipartition of one cluster at each step. In this paper we address divisive hierarchical clustering, and extend the divisive algorithm proposed in (Chavent, 1998) and (Chavent et al., 2007) to data described by interval and/or histogram-valued variables. The method successively splits one cluster into two sub-clusters, according to a condition expressed as a binary question on the values of one variable; the cluster to be split and the condition to be considered at each step are selected so as to minimize intra-cluster dispersion on the next step. Therefore, each formed cluster is automatically interpreted by a conjunction of necessary and sufficient conditions for cluster membership (the conditions that lead to its formation by successive splits) and we obtain a monothetic clustering (Sneath and Sokal, 1973) on the dataset.

There is a variety of divisive clustering methods (Kaufman and Rousseeuw, 1990). A natural approach of dividing a cluster C of n objects into two non-empty subsets would be to consider all the possible bipartitions; however, a complete enumeration procedure provides a global optimum but is computationally prohibitive. Nevertheless, it is possible to construct divisive clustering methods that do not consider all bipartitions. In (MacNaughton-Smith, 1964) an iterative divisive procedure is proposed that uses an average dissimilarity between an object and a group of objects; (Gowda and Krishna, 1978) proposed a disaggregative clustering method based on the concept of mutual nearest neighborhood. Monothetic divisive clustering methods have first been proposed for binary data (Williams and Lambert, 1959), (Lance and Williams, 1968); then, monothetic clustering methods were mostly developed for unsupervised learning and are known as descendant conceptual clustering methods (Michalski et al., 1981), (Michalski and Stepp, 1983). Other approaches may be referred in the context of information-theoretic clustering (Dhillon et al., 2003) and spectral clustering (Boley, 1998), (Fang and Saad, 2008). In the field of discriminant analysis, monothetic divisive methods have also been widely developed: a partition is pre-defined and the problem concerns the construction of a systematic way of predicting the class membership of a new object. In Pattern Recognition literature, this type of classification is referred to as supervised pattern recognition. Di-

visive methods of this type are usually known as tree structured classifier like CART (Breiman et al., 1984) or ID3 (Quinlan, 1986). In (Ciampi, 1994) the author stresses the idea that trees offer a natural approach for both class formation (clustering) and development of classification rules.

In classical statistics and multivariate data analysis, the basic units under analysis are single individuals, described by numerical and/or categorical variables, each individual taking one single value for each variable. For instance, a specific man may be described by his age, weight, color of the eyes, etc. Data are organized in a data-array, where each cell (i, j) contains the value of variable j for individual i. This model is however too restricted to take into account variability and/or uncertainty which are often inherent to the data. When analyzing a group rather than a single individual, then variability intrinsic to the group should be taken into account. Consider, for instance, that we are analyzing the staff of some given institutions, in terms of age, marital status and category. If we just take averages or mode values within each institution, much information is lost. Also, when we observe some given variables along time, and wish to record the set of observed values rather than a single statistics (e.g., mean, maximum,...), then again a set of values rather than a single one must be recorded. The same issue arises when we are interested in concepts and not in single specimen - whether it is a plant species (and not the specific plant I have in my hand), a model of car (and not the one I am driving), etc. Whether the data are obtained by contemporaneous or temporal aggregation of individual observations to obtain descriptions of the entities which are of interest, or whether we are facing concepts as such specified by experts or put in evidence by clustering, we are dealing with elements which can no longer be properly described by the usual numerical and categorical variables without an unacceptable loss of information. Symbolic Data Analysis - see (Bock and Diday, 2000), (Billard and Diday, 2006), (Diday and Noirhomme-Fraiture, 2008) or (Noirhomme-Fraiture and Brito, 2011) - provides a framework where the variability observed may effectively be considered in the data representation, and methods be developed that take it into account. To describe groups of individuals or concepts, variables may now assume other forms of realizations, which allow taking into account the intrinsic variability. These new variable types have been called "symbolic variables", and they may assume multiple, possibly weighted, values for each entity. Data are gathered in a matrix, now called a "symbolic data table", each cell containing "symbolic data". To each row of the table corresponds a group, or concept, i.e., the entity of interest. A numerical variable may then be single valued (real or integer), as in the classical framework, if it takes one single value of an underlying domain per entity, it is multi-valued if its values are finite subsets of the domain and it is an interval variable if its values are intervals. When an empirical distribution over a set of sub-intervals is given, the variable is called a histogram-valued variable - see (Bock and Diday, 2000) and (Noirhomme-Fraiture and Brito, 2011).

Several clustering methods for symbolic data have been developed. The divisive clustering algorithm, proposed in (Chavent, 1998) and (Chavent et al., 2007), has been extended to the case of interval-valued variables and modal categorical variables (i.e., variables for which a distribution on a finite set of categories is observed), see (Chavent, 2000). This is however a different approach to the one proposed here, in that it does not allow for mixed variable types, no order is considered in the category set (whereas for histogram-valued variables the considered sub-intervals are naturally ordered), and the distances allowing to evaluate intra-cluster dispersion are not the same. Extensions of the *k-means* algorithm, may be found, for instance, in (De Souza and De Carvalho, 2004), (De Carvalho et al., 2006), (Chavent et al., 2006), (De Carvalho et al., 2009) and (De Carvalho and De Souza, 2010). A method based on Poisson point processes has been proposed in (Hardy and Kasaro, 2009); clustering and validation of interval data are discussed in (Hardy and Baune, 2007). A method for "symbolic" hierarchical or pyramidal clustering has been proposed in (Brito, 1994) and (Brito, 1995), which allows clustering multi-valued data of different types; it was subsequently developed in order to allow for variables for which distributions on a finite set are recorded (Brito, 1998). It is a conceptual clustering method, since each cluster formed is associated with a conjunction of properties in the input variables, which constitutes a necessary and sufficient condition for cluster membership. On a recent approach, (Irpino and Verde, 2006) propose using the Wasserstein distance for clustering histogram-valued data. For more details on clustering for symbolic data see (Billard and Diday, 2006) and (Diday and Noirhomme-Fraiture, 2008); in (Noirhomme-Fraiture and Brito, 2011) an extensive survey is presented.

The remaining of the paper is organized as follows. Section 2 presents interval and histogram-valued variables, introducing the new types of realizations. In Section 3 the proposed clustering method is described, detailing the different options to be made. An illustrative example is presented in 4. Section 5 concludes the paper, pointing paths for further research.

2 INTERVAL AND HISTOGRAM-VALUED DATA

Let $\Omega = \{\omega_1, ..., \omega_n\}$ be the set of n objects to be analyzed, and $Y_1, \ldots, Y_p$ the descriptive variables.

2.1 Interval-valued variables

An interval-valued variable is defined by an application $Y_j : \Omega \to B$ such that $Y_j(\omega_i) = [l_{ij}, u_{ij}], l_{ij} \leq u_{ij}$, where B is the set of intervals of an underlying set $O \subseteq \mathbb{R}$. Let I be an $n \times p$ matrix representing the values of p interval variables on Ω. Each $\omega_i \in \Omega$ is represented by a p-tuple of intervals, $I_i = (I_{i1}, ..., I_{ip}), i = 1, ..., n$, with $I_{ij} = [l_{ij}, u_{ij}], j = 1, \ldots, p$ (see Table 1).

Table 1: Matrix I of interval data.

	Y_1	$\ldots$	Y_j	$\ldots$	Y_p
ω_1	$[l_{11}, u_{11}]$	$\ldots$	$[l_{1j}, u_{1j}]$	$\ldots$	$[l_{1p}, u_{1p}]$
$\ldots$	$\ldots$		$\ldots$		$\ldots$
ω_i	$[l_{i1}, u_{i1}]$	$\ldots$	$[l_{ij}, u_{ij}]$	$\ldots$	$[l_{ip}, u_{ip}]$
$\ldots$	$\ldots$		$\ldots$		$\ldots$
ω_n	$[l_{n1}, u_{n1}]$	$\ldots$	$[l_{nj}, u_{nj}]$	$\ldots$	$[l_{np}, u_{np}]$

Example 1
Consider three persons, Albert, Barbara and Caroline characterized by the amount of time (in minutes) they need to go to work, which varies from day to day and is therefore represented by an interval-valued variable, as presented in Table 2.

Table 2: Amount of time (in minutes) necessary to go to work for three persons.

	Time
Albert	$[15, 20]$
Barbara	$[25, 30]$
Caroline	$[10, 20]$

2.2 Histogram-valued variables

A histogram variable Y_j is defined by an application $Y_j : \Omega \to B$ where B is now the set of probability or frequency distributions in the considered sub-intervals $\{I_{ij1}, ..., I_{ijk_{ij}}\}$ (k_{ij} is the number of sub-intervals in $Y_j(\omega_i)$); $Y_j(\omega_i) = (I_{ij1}, p_{ij1}; \ldots; I_{ijk_{ij}}, p_{ijk_{ij}})$ with $p_{ij\ell}$ the probability or frequency associated to the sub-interval $I_{ij\ell} = [\underline{I}_{ij\ell}, \overline{I}_{ij\ell}[$ and $p_{ij1} + \ldots + p_{ijk_{ij}} = 1$.
Therefore, $Y_j(\omega_i)$ may be represented by the histogram (Bock and Diday, 2000):

$$H_{Y_j(\omega_i)} = ([\underline{I}_{ij1}, \overline{I}_{ij1}[, p_{ij1}; \ldots; [\underline{I}_{ijk_{ij}}, \overline{I}_{ijk_{ij}}], p_{ijk_{ij}}) \qquad (1)$$

$i \in \{1, 2, \ldots, n\}, \underline{I}_{ij\ell} \leq \overline{I}_{ij\ell}$ and $\overline{I}_{ij\ell} \leq \underline{I}_{ij(\ell+1)}$.

It is assumed that within each sub-interval $[\underline{I}_{ij\ell}, \overline{I}_{ij\ell}[$ the values of variable Y_j for observation ω_i, are uniformly distributed. For each variable Y_j the number and length of sub-intervals in $Y_j(\omega_i)$, $i = 1, \ldots, n$ may naturally be different. To apply a clustering method, however, all observations of each histogram-valued variable should be written using the same underlying partition, so that they are directly comparable. For each variable, we then re-write each observed histogram using the intersection of the given partitions, and assuming a uniform distribution within each sub-interval. Example 2 below illustrates the procedure. Once the observed histograms have been re-written, each histogram-valued variable is written on the same partition, let now K_j be the number of sub-intervals for variable $j, j = 1, \ldots, p$.
For each observation ω_i, $Y_j(\omega_i)$ can, alternatively, be represented by the inverse cumulative distribution function, also called quantile function, q_{ij} - see (Irpino and Verde, 2006) - given by

$$q_{ij}(t) = \begin{cases} \underline{I}_{ij1} + \frac{t}{w_{i1}} r_{ij1} \; if \; 0 \leq t < w_{i1} \\ \underline{I}_{ij2} + \frac{t - w_{i1}}{w_{i2} - w_{i1}} r_{ij2} \; if \; w_{i1} \leq t < w_{i2} \\ \vdots \\ \underline{I}_{ijK_j} + \frac{t - w_{iK_j - 1}}{1 - w_{ijK_j - 1}} r_{ijK_j} \; if \; w_{ijK_{ij}-1} \leq t \leq 1 \end{cases}$$

where $\quad w_{ih} = \sum_{\ell=1}^{h} p_{ij\ell}, h = 1, \ldots, K_j; r_{ij\ell} = \overline{I}_{ij\ell} - \underline{I}_{ij\ell}$
for $\ell = \{1, \ldots, K_j\}$.

Notice that interval-valued variables may be considered as a particular case of histogram-valued variables, where $Y_j(\omega_i) = [l_{ij}, u_{ij}]$ may be written as $H_{Y_j(\omega_i)} = ([l_{ij}, u_{ij}], 1)$. In this case, rather than re-writing each observation using the same partition, they must be re-written for the same weight distribution, to allow for the comparision of the corresponding quantile functions.

Example 2
Consider now two classes of students, for which the age range and the distribution of the marks obtained in an exam were registered, as presented in Table 3. Students in Class 1 have ages ranging from 10 to 12 years old, 20% of them had marks between 5 and 10, 50% had marks between 10 and 12, 20% between 12 and 15 and 10% between 15 and 18; likewise for Class 2. Notice that the units of interest here are the classes as a whole and not each individual student.
In Table 4, the values of the histogram-valued variable "Marks" are re-written on the intersection partitons.

Figure 1 represents the the observed histograms of variable "Marks" for Class 1 and Class 2. Figure 2 de-

Table 3: Age range and distribution of obtained marks for two classes of students.

	Age	Marks
Class 1	$[10,12]$	$([5,10[,0.2;[10,12[,0.5;$ $[12,15[,0.2;[15,18],0.1)$
Class 2	$[11,14]$	$([5,10[,0.05;[10,12[,0.3;[12,14[,0.25;$ $[14,16[,0.2;[16,19],0.2;)$

Table 4: Age range and distribution of obtained marks for two classes of students, after re-writing the histograms with the intersection partitions.

	Age	Marks
Class 1	$([10,11[,0.5;$ $[11,12[,0.5;$ $[12,14],0.0)$	$([5,10[,0.2;[10,12[,0.5;[12,14[,0.133;$ $[14,15[,0.067;[15,16[,0.033;$ $[16,18[,0.067;[18,19],0.0)$
Class 2	$([10,11[,0;$ $[11,12[,0.33;$ $[12,14],0.67)$	$([5,10[,0.05;[10,12[,0.3;[12,14[,0.25;$ $[14,15[,0.1;[15,16[,0.1;$ $[16,18[,0.133;[18,19],0.067)$

picts the respective quantile functions, obtained after re-writing the histograms with the same partition.

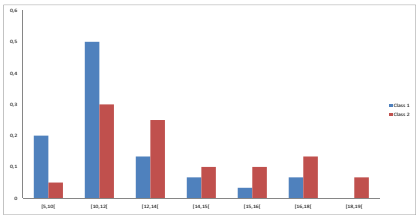

Figure 1: Representation of Marks(Class 1) and Marks(Class 2) in the form of histograms.

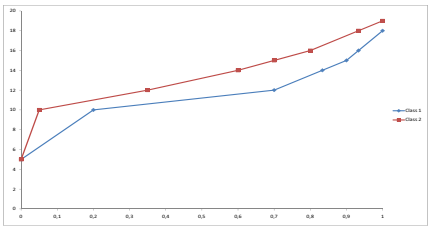

Figure 2: Representation of Marks(Class 1) and Marks(Class 2) by the quantile functions.

Henceforth "distribution" refers to a probability or frequency distribution of a continuous variable represented by a histogram or a quantile function.

3 DIVISIVE CLUSTERING

Divisive clustering algorithms proceed top-down, starting with Ω, the set to be clustered, and performing a bipartition of one cluster at each step. At step m a partition of Ω in m clusters is present, one of which will be further divided in two sub-clusters; the cluster to be divided and the splitting rule are chosen so as to obtain a partition in $m+1$ clusters minimizing intra-cluster dispersion.

3.1 The criterion

The "quality" of a given partition $P_m = \left\{C_1^{(m)}, C_2^{(m)}, \ldots, C_m^{(m)}\right\}$ is measured by a criterion $Q(m)$, the sum of intra-cluster dispersion for each cluster :

$$Q(m) = \sum_{\alpha=1}^{K} I(C_\alpha) = \sum_{\alpha=1}^{K} \sum_{\omega_i,\omega_{i'} \in C_\alpha^{(m)}} D^2(\omega_i,\omega_{i'}) \quad (2)$$

$$\text{with } D^2(\omega_i,\omega_{i'}) = \sum_{j=1}^{p} d^2(x_{ij},x_{i'j}) \quad (3)$$

where d is a quadratic distance between distributions (notice that both for interval-valued and histogram-valued variables, x_{ij} is represented as a distribution). That is, for each cluster, intra-cluster dispersion is defined as the sum D^2 of all pairwise squared-distances between the cluster elements. We consider distances D^2 additive on the descriptive variables.
At each step one cluster is chosen to be split in two sub-clusters, so that $Q(m+1)$ is minimized, or, equivalently, $Q(m)-Q(m+1)$ maximized (notice that Q always decreases at each step).

3.1.1 Distances

Several distances may be considered to evaluate the dissimilarity between distributions. Let $Y_j(\omega_i) = H_{Y_j(\omega_i)} = ([\underline{I}_{ij1},\overline{I}_{ij1}[,p_{ij1};\ldots;[\underline{I}_{ijK_j},\overline{I}_{ijK_j}],p_{ijK_j})$. We propose to use one of the two following distances:

1. Mallows distance
$$d_M^2(x_{ij},x_{i'j}) = \int_0^1 (q_{ij}(t)-q_{i'j}(t))^2 dt$$
q_{ij} is the quantile function corresponding to the distribution $Y_j(\omega_i)$.
2. Squared Euclidean distance
$$d_E^2(x_{ij},x_{i'j}) = \sum_{\ell=1}^{K_j} (p_{ij\ell}-p_{i'j\ell})^2$$

The Mallows distance has been used in agglomerative hierarchical clustering in (Irpino and Verde, 2006).

3.2 Binary Questions and Assignment

The bipartition to be performed at each step is defined by one single variable, considering conditions of the

type $R_{j\ell} := Y_j \leq \bar{I}_{j\ell}, \ell = 1,\ldots,K_j - 1, j = 1,\ldots,p$, i.e., we consider the upper bounds $\bar{I}_{j\ell}$ of all sub-intervals (except for the last one) corresponding to each variable.
Each condition $R_{j\ell}$ leads to a bipartition of a cluster, sub-cluster 1 gathers the elements who verify the condition, sub-cluster 2 those who do not. An element $\omega_i \in \Omega$ verifies the condition $R_{j\ell} = Y_j \leq \bar{I}_{j\ell}$ iff $\sum_{\alpha=1}^{\ell} p_{ij\alpha} \geq 0.5$ (Chavent, 2000).
Notice that the sequence of conditions met by the elements of each cluster constitutes a necessary and sufficient condition for cluster membership. The obtained clustering is therefore monothetic, i.e. each cluster is represented by a conjunction of properties in the descriptive variables.

Example 3
Consider again the data in Example 2. At the first step, the binary questions to be considered are :
Age $\leq$ 11, Age $\leq$ 12, Marks $\leq$ 10, Marks $\leq$ 12, Marks $\leq$ 14, Marks $\leq$ 15, Marks $\leq$ 16, Marks $\leq$ 18.
If condition Age $\leq$ 12 is selected, then sub-cluster 1 shall contain Class 1, and be described by "Age $\leq$ 12", and sub-cluster 2 shall contain Class 2 and be described by "Age $>$ 12".

At each step, the cluster $C_\ell^{(m)}$ and the splitting condition $R_{j\ell}$ are chosen so that the resultant partition P_{m+1}, in $m+1$ clusters) minimizes $Q(m+1)$.

3.3 The Algorithm

The proposed divisive clustering algorithm may now be summarized as follows. Let $P_m = \{C_1^{(m)},\ldots,C_m^{(m)}\}$ be the current partition at step m.
Initialization : $P_1 = \{C_1^{(1)} \equiv \Omega\}$. At step m: Determine the cluster $C_M^{(m)}$ and the binary question $R_{j\ell} := Y_j \leq \bar{I}_{j\ell}, \ell = 1,\ldots,K_j, j = 1,\ldots,p$, such that the new resulting partition $P_{m+1} = \{C_1^{(m+1)},\ldots,C_{m+1}^{(m+1)}\}$, in $m+1$ clusters, minimizes intra-cluster dispersion, given by $Q(m) = \sum_{\ell=1}^{m} \sum_{\omega_i,\omega_{i'} \in C_\ell^{(m)}} \sum_{j=1}^{p} d^2(x_{ij}, x_{i'j}))$, among partitions in $m+1$ clusters obtained by splitting a cluster of P_m in two clusters. Notice that to minimize Q(m) is equivalent to maximize $\Delta Q = I(C_M^{(m)}) - (I(C_1^{(m+1)}) + I(C_2^{(m+1)}))$.
When the desired, pre-fixed, number of clusters is attained, or P has n clusters, each with a single element (step n), the algorithm stops.

4 ILLUSTRATIVE EXAMPLE

Table 5 gathers information about the Price (in thousands of dollars) and Engine Displacement (in cm^3) of four utilitarian cars' models, considering histograms, already written with the same partitions.

Table 5: Price and Engine Displacement of 4 car models.

	Price	Engine Displacement
Model 1	$([15,25[,0.5;[25,35[,0.5)$;	$([1300,1500[,0.2;[1500,1700[,0.5;$ $[1700,1900[,0.3)$
Model 2	$([15,25[,0.2;[25,35[,0.8)$;	$([1300,1500[,0.1;[1500,1700[,0.2;$ $[1700,1900[,0.7)$
Model 3	$([15,25[,0,33;[25,35[,0.67)$	$([1300,1500[,0.1;[1500,1700[,0.4;$ $[1700,1900[,0.5)$
Model 4	$([15,25[,0.6;[25,35[,0.4)$	$([1300,1500[,0.6;[1500,1700[,0.4;$ $[1700,1900[,0.0)$

A partition into three clusters is desired. We choose to use the squared Euclidean distance between distributions to compare the observed values for each car model. The distance matrice is then

$$D = \begin{pmatrix} 0.0000 & 0.4400 & 0.1178 & 0.2800 \\ 0.4400 & 0.0000 & 0.1380 & 1.1000 \\ 0.1178 & 0.1380 & 0.0000 & 0.8429 \\ 0.2800 & 1.1000 & 0.8429 & 0.0000 \end{pmatrix}$$

We start with the trivial partition (in one cluster) $P_1 = \{\Omega = \{\text{Model 1 , Model 2 , Model 3 , Model 4}\}\}$. At step 2, Ω is split into clusters $C_1^{(2)} = \{\text{Model 1, Model 4}\}$ and $C_2^{(2)} = \{\text{Model 2, Model 3}\}$, according to condition $R_{11} :=$ Price ≤ 25; partition $P_2 = \{C_1^{(2)}, C_2^{(2)}\}$ has total intra-cluster dispersion equal to 0.3938. At step 3, $C_1^{(2)}$ is further divided into $C_1^{(3)} = \{$ Model 4 $\}$ and $C^(3)_2 = \{$ Model 1 $\}$, according to condition $R_{22} :=$ Engine Displacement ≤ 1500. A partition in three clusters $P_3 = \{C_1^{(3)}, C_2^{(3)}, C_3^{(3)} = C_2^{(2)}\} = \{\{\text{Model 4}\}, \{\text{Model 1}\}, \{\text{Model 2, Model 3}\}\}$ is obtained, with intra-cluster dispersion equal to 0.1132. Cluster $C_1^{(3)} = \{\text{Model 4}\}$ is described by "Price $\leq$ 25 $\wedge$ Engine Displacement $\leq$ 1500"; Cluster $C_2^{(3)} = \{\text{Model 1}\}$ by "Price $\leq$ 25 $\wedge$ Engine Displacement $>$ 1500"; Cluster $C_3^{(3)} = \{\text{Model 2, Model 3}\}$ by "Price $>$ 25".

5 CONCLUSIONS

We have proposed a divisive clustering method for data described by interval and/or histogram-valued variables. The method provides a hierarchy on the set under analysis, together with a conjunctive characterization of each cluster. Distances for comparing distributions are considered. Experiments with real-data allowing for comparision with alternative meth-

ods are planned.

The following step should consist of implementing a procedure for revising the condition inducing the cluster chosen for splitting at each step, so as to improve the obtained clustering (in the same line as in (Chavent, 1998)). Also, the hierarchy may be indexed, so that a dendrogram is obtained. Finally, the complexity of the computation of the intra-cluster dispersion may be reduced by taking into account the order between the cut values $\overline{I}_{j\ell}$. These developments will be the subject of our further research.

REFERENCES

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley.

Bock, H.-H. and Diday, E. (2000). *Analysis of Symbolic Data*. Springer, Berlin-Heidelberg.

Boley, D. L. (1998). Principal direction divisive partitioning. *Data Mining and Knowledge Discovery*, 2(4):325–344.

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

Brito, P. (1994). Use of pyramids in symbolic data analysis. In Diday, E. et al., editors, *New Approaches in Classification and Data Analysis*, pages 378–386, Berlin-Heidelberg. Springer.

Brito, P. (1995). Symbolic objects: order structure and pyramidal clustering. *Annals of Operations Research*, 55:277–297.

Brito, P. (1998). Symbolic clustering of probabilistic data. In Rizzi, A. et al., editors, *Advances in Data Science and Classification*, pages 385–389, Berlin-Heidelberg. Springer.

Chavent, M. (1998). A monothetic clustering method. *Pattern Recognition Letters*, 19(11):989–996.

Chavent, M. (2000). Criterion-based divisive clustering for symbolic objects. In Bock, H.-H. and Diday, E., editors, *Analysis of Symbolic Data*, pages 299–311, Berlin-Heidelberg. Springer.

Chavent, M., De Carvalho, F. A. T., Lechevallier, Y., and Verde, R. (2006). New clustering methods for interval data. *Computational Statistics*, 21(2):211–229.

Chavent, M., Lechevallier, Y., and Briant, O. (2007). DIVCLUS-T: A monothetic divisive hierarchical clustering method. *CSDA*, 52(2):687–701.

Ciampi, A. (1994). Classification and discrimination: the RECPAM approach. In Dutter, R. and Grossmann, W., editors, *Proc. COMPSTAT'94*, pages 129–147, Heidelberg. Physica Verlag.

De Carvalho, F. A. T., Brito, P., and Bock, H.-H. (2006). Dynamic clustering for interval data based on L_2 distance. *Computational Statistics*, 21(2):231–250.

De Carvalho, F. A. T., Csernel, M., and Lechevallier, Y. (2009). Clustering constrained symbolic data. *Pattern Recognition Letters*, 30(11):1037–1045.

De Carvalho, F. A. T. and De Souza, R. M. C. R. (2010). Unsupervised pattern recognition models for mixed feature-type symbolic data. *Pattern Recognition Letters*, 31(5):430–443.

De Souza, R. M. C. R. and De Carvalho, F. A. T. (2004). Clustering of interval data based on city-block distances. *Pattern Recognition Letters*, 25(3):353–365.

Dhillon, I. S., Mallela, S., and Kumar, R. (2003). A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, (3):1265–1287.

Diday, E. and Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the Sodas Software*. Wiley.

Fang, H. and Saad, Y. (2008). Farthest centroids divisive clustering. In *Proc. ICMLA*, pages 232–238.

Gowda, K. C. and Krishna, G. (1978). Disaggregative clustering using the concept of mutual nearest neighborhood. *IEEE Trans. SMC*, 8:888–895.

Hardy, A. and Baune, J. (2007). Clustering and validation of interval data. In Brito, P. et al., editors, *Selected Contributions in Data Analysis and Classification*, pages 69–82, Heidelberg. Springer.

Hardy, A. and Kasaro, N. (2009). A new clustering method for interval data. *MSH/MSS*, 187:79–91.

Irpino, A. and Verde, R. (2006). A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In Batagelj, V. et al., editors, *Proc. IFCS 2006*, pages 185–192, Heidelberg. Springer.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data*. Wiley, New York.

Lance, G. N. and Williams, W. T. (1968). Note on a new information statistic classification program. *The Computer Journal*, 11:195–197.

MacNaughton-Smith, P. (1964). Dissimilarity analysis: A new technique of hierarchical subdivision. *Nature*, 202:1034–1035.

Michalski, R. S., Diday, E., and Stepp, R. (1981). A recent advance in data analysis: Clustering objects into classes characterized by conjunctive concepts. In Kanal, L. N. and Rosenfeld, A., editors, *Progress in Pattern Recognition*, pages 33–56, Amsterdam. North-Holland.

Michalski, R. S. and Stepp, R. (1983). Learning from observations: Conceptual clustering. In Michalsky, R. S. et al., editors, *Machine Learning: An Artificial Intelligence Approach*, pages 163–190. Morgan Kaufmann.

Noirhomme-Fraiture, M. and Brito, P. (2011). Far beyond the classical data models: Symbolic data analysis. *Statistical Analysis and Data Mining*, 4(2):157–170.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1:81–106.

Sneath, P. H. and Sokal, R. R. (1973). *Numerical Taxonomy*. Freeman, San Francisco.

Williams, W. T. and Lambert, J. M. (1959). Multivariate methods in plant ecology. *J. Ecology*, 47:83–101.