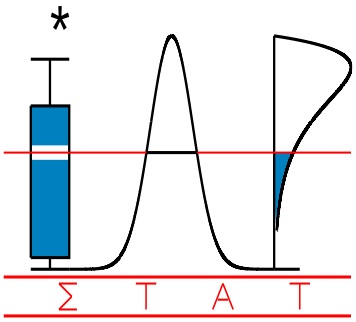
**T E C H N I C A L**

**R E P O R T**

**0437**

# DYNAMIC CLUSTERING FOR INTERVAL DATA BASED ON $L_2$ DISTANCE

DE CARVALHO, F., BRITO, P. and H.-H. BOCK




**I A P S T A T I S T I C S**

**N E T W O R K**

**INTERUNIVERSITY ATTRACTION POLE**

http://www.stat.ucl.ac.be/IAP

# Preprint from 11/6/2004

# Dynamic Clustering for Interval Data Based on $L_2$ Distance

Francisco de A. T. de Carvalho [1,*]

*Centro de Informatica - CIn/UFPE, Av. Prof Luiz Freire, s/n, Cidade Universitaria, CEP 50.740-540, Recife - PE BRAZIL*

Paula Brito

*Faculdade de Economia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-464 Porto, PORTUGAL*

Hans-Hermann Bock

*RWTH Aachen University, Institute of Statistics, 52056 Aachen, GERMANY*

**Abstract**

This paper introduces a partitioning clustering method for objects described by interval data. It follows the dynamic clustering approach and uses an $L_2$ distance. Particular emphasis is put on the standardization problem where we propose and investigate three standardization techniques for interval-type variables. Moreover, various tools for cluster interpretation are presented and illustrated by simulated and real-case data.

*Key words:* Clustering, Symbolic Data Analysis, Interval Data, Standardization, Cluster Interpretation.

* Corresponding Author. Tel.:+55-81-21268430; fax:+55-81-21268438
*Email addresses:* fatc@cin.ufpe.br (Francisco de A. T. de Carvalho ), mpbrito@fep.up.pt (Paula Brito), bock@stochastik.rwth-aachen.de (Hans-Hermann Bock).

[1] Acknowledgements. The first author would like to thank CNPq (Brazilian Agency) for its financial support. The last author gratefully acknowledges the support received by the IAP research network no. P5/24 of the Belgian State (Federal Office for Scientific, Technical and Cultural Affairs).

## 1 Introduction and survey on clustering for symbolic data

Clustering is a multivariate statistical technique that aims at collecting similar objects in homogeneous clusters, on the basis of observed values for a set of variables. The constructed clusters may be organized according to different set-theoretical structures: partitioning methods construct a partition of the set of objects with non-overlapping clusters, usually optimizing some suitable criterion, whereas hierarchical and pyramidal clustering methods produce a system of nested and stratified clusters. For a good overview on cluster analysis see, for instance, Jain *et al.* (1999) or Gordon (1999).

In this paper, we present a partitioning clustering method based on the dynamic clustering methodology. Dynamical clustering (Diday (1972), Diday and Simon (1976)), also known as $k$-means, alternating minimization etc., provides a general framework for non-hierarchical clustering which has given rise to many different particular methods for special clustering methods and data cases. The method allows to obtain a partition of a set $\Omega = \{1, ..., n\}$ of objects into a given number $k$ of clusters, by optimizing a criterion that evaluates the fit between the cluster members and their representatives. The general algorithm proceeds by an iterative application of a *class assignment function* that allows to form clusters, and a corresponding *representation function* that determines the class representatives (prototypes) of the formed clusters, until convergence is attained. Various special algorithms result from different choices of the dissimilarity measure used by the assignment function, and of the representation space: centroids (Diday (1972)), factorial axes (Bock (1974), Ok-Sakun (1975)), probability laws (Schroeder (1976)), etc.

In the classical data table model each row represents an individual (or object) and each column represents a variable. In particular, each individual takes just one single value for each variable. In practice, however, many situations do not fit this simple model since there are several values or categories for each variable, possibly even with frequencies or weights. Such data, together with the idea of producing results which are directly interpretable in terms of the input variables, have led to the development of *Symbolic Data Analysis*. Thereby new types of variables - interval-type, categorical multi-valued, and modal variables - have been introduced which allow to represent the variability and/or uncertainty that underly the observed data (Bock and Diday (2000)). The main objective of Symbolic Data Analysis is to extend classical data analysis techniques to such 'symbolic' data whereby the input data and the output results may both be expressed within the same formalism, based on the notion of a 'symbolic object'.

As far as clustering is concerned, the problem consists in developing methods that allow to cluster a set of individuals described by symbolic data and to

produce classes which are directly interpretable in terms of the input variables. Nowadays, many clustering methods for symbolic data have been proposed which differ in the type of the considered symbolic variables, in their cluster structures and/or in the considered clustering criteria (see Bock and Diday (2000), Bock (2004)).

Several authors have addressed the problem of non-hierarchical clustering for symbolic data. Diday and Brito (1989) used a transfer algorithm to partition a set of symbolic objects into clusters described by weight distribution vectors. Ralambondrainy (1995) extended the classical $k$-means clustering method in order to deal with data characterized by numerical and categorical variables, and complemented his method by a characterization algorithm to provide a conceptual interpretation of the resulting clusters. El-Sonbaty and Ismail (1998) have presented a fuzzy $k$-means algorithm to cluster data on the basis of different types of symbolic variables. Gordon (2000) presented an iterative relocation algorithm to partition a set of symbolic objects into classes so as to minimize the sum of the description potentials of the classes. Verde *et al.* (2001) introduced a dynamic clustering algorithm for symbolic data considering context-dependent proximity functions where the cluster representatives are weight distribution vectors. Bock (2002) has proposed several clustering algorithms for symbolic data described by interval variables, based on a clustering criterion and thereby generalized similar approaches in classical data analysis. Chavent and Lechevallier (2002) proposed a dynamic clustering algorithm for interval data where the class representatives are defined by an optimality criterion based on a modified Hausdorff distance. More recently, Souza and De Carvalho (2004) have proposed partitioning clustering methods for interval data based on city-block distances, also considering adaptive distances.

In this paper, we address the problem of clustering interval data. A partitioning clustering method is presented which allows to cluster objects described both by classical quantitative and interval-type variables. It is based on the dynamic clustering approach and uses a Minkowski-like distance of $L_2$ type (Ichino and Yaguchi (1994)) which results as a special case of the distance introduced by De Carvalho and Souza (1998). Each cluster is represented by a class prototype that is a vector of intervals such that, for each coordinate or variable, the corresponding interval has, as its lower bound, the average of all lower bounds of the intervals of all cluster members, and the average of the upper bounds of these intervals as its upper bound. This is a 'most typical' class prototype under a vertex-type dissimilarity measure for interval vectors Bock (2004).

When clustering quantitative data, standardization is an important issue Indeed, the result of any clustering method depends heavily on the scales used for the variables: modifying the scales of single variables may disturb or de-

stroy a natural clustering whereas, inversely, natural clustering structures can sometimes only be detected after an appropriate rescaling of variables before running a clustering algorithm (e.g., by transforming all variables to the same standard deviation). In this paper, we address the standardization problem for interval variables and propose three alternative standardization techniques. The two first ones are based on the usual mean-and-variance paradigm, but differ in the way of measuring the dispersion of a set of intervals: the first one measures the overall dispersion by the dispersion of the *interval centers* whereas the second one uses the dispersion of the *interval boundaries*. The third method transforms the interval variables such that the resulting global range becomes the full unit interval $[0, 1]$.

After having obtained a clustering of the underlying objects by an algorithm, it is important to interpret and evaluate its classes. Celeux *et al.* (1989) have introduced a family of indices to interpret a partition of classical quantitative data, based on the notion of 'inertia'. In this paper, we adapt these indices to the case of interval data and propose various indices for evaluating the quality of a partition, the homogeneity and eccentricity of the individual clusters, and the role played by the different variables in the cluster formation process. Typically, all resulting clusters may be represented by symbolic objects which provide a conceptual description directly in terms of the observed variables. We will consider two alternative descriptions: a first one by using the cluster representatives, and a second one which provides a 'generalizing description' of the cluster. By computing the degree of overlap of the symbolic descriptions of pairs of clusters, we may then evaluate their mutual separation.

The proposed methods have been tested on simulated and real data. Two simulated data sets have been generated, one with well separated clusters and another one with overlapping clusters, so as to study the performance of the method under three alternative standardizations. Our evaluation of the clustering results is based on an external validity index, i.e., the corrected Rand index (Hubert and Arabie (1985)), in the framework of a Monte Carlo study with 100 replications of each data set. The mean and standard deviation of the values of the external validity index are computed for each method. Additionally, our method has been applied to a real data-set consisting of 33 car models described by 8 interval variables. In this study we analyze the role of the different variables in the clustering process and stress the need for standardization. Moreover, we explore the potentialities of the proposed family of interpretation indices.

Section 2 recalls the dynamic clustering approach. Section 3 introduces symbolic data and specifies the data model considered in our method. Section 4 presents our clustering algorithm and discusses the influence of the selected distance measure. Section 5 addresses the standardization issue. In Section 6, we propose tools for the interpretation of the obtained partition. Section

7 presents and discusses the results of our numerical experiments. Finally, Section 8 concludes and opens perspectives for further developments.

## 2 Dynamic Clustering

Dynamic clustering (Diday and Simon (1976)) is a non-hierarchical clustering method that determines a partition of a set $\Omega = \{1, ..., n\}$ of objects into a fixed number $k$ of clusters by optimizing a criterion $W$ that evaluates the fit between the clusters and their representatives. The method requires the suitable definition of a 'representation' of a given set of objects (a cluster representative or prototype) which may be a central object, a subset of objects, an interval, a function, etc. Starting from a set of class representatives or from an initial partition (random or user defined), the method applies an *assignment function* followed by a cluster *representation function* and iterates these steps in turn until convergence is achieved.

In the following we consider a suitable set $\mathcal{L}$ of 'admissible' class representatives (the representation space) and denote by $\mathcal{L}_k = (\mathcal{L})^k$ the set of all k-tuples of admissible representatives $L = (\ell_1, \ldots, \ell_k)$ (one for each cluster). Let $\mathcal{P}_k$ denote the family of all partitions $P = (P_1, \ldots, P_k)$ of $\Omega$ into the given number $k$ of non-empty clusters, and $D(\ell_h, P_h)$ a dissimilarity measure between a class representative $\ell_h$ and a cluster $P_h$ (a low value indicates a good fit between $\ell_h$ and $P_h$). In this paper, we consider a clustering criterion of the form

$$W(P, L) = \sum_{h=1}^{k} D(\ell_h, P_h) \tag{1}$$

which can be considered as a mapping $W : \mathcal{P}_k \times \mathcal{L}_k \longrightarrow \Re^+$. Our clustering problem consists in finding a pair $(L^*, P^*) \in \mathcal{P}_k \times \mathcal{L}_k$ that minimizes W, i.e., such that

$$W(P^*, L^*) = min\{W(P, L) : P \in \mathcal{P}_k, L \in \mathcal{L}_k\}. \tag{2}$$

In this paper the dissimilarity (heterogeneity) measure $D(\ell, P)$ is obtained by considering a dissimilarity measure $\varphi(x, y)$ for elements $x, y$ of the representation space $\mathcal{L}$ (later on this will be the set $\mathcal{I}^p$ of all $p$-dimensional rectangles of $\Re^p$) and summing up over all individuals $\omega_i$ of the class $P$ (with corresponding data rectangles $x_i \in \mathcal{L}$):

$$D(\ell, P) := \sum_{\omega_i \in P} \varphi(\ell, x_i) \qquad \ell \in \mathcal{L}, P \subset \Omega. \tag{3}$$

A dynamic clustering (or $k$-means) algorithm is defined in terms of an *assignment function* $f$ and a *representation function* $g$. They are defined as follows:

*The assignment function* $f : \mathcal{L}_k \longrightarrow \mathcal{P}_k$ is the function that assigns to each $k$-tuple $L = (\ell_1, \ldots, \ell_k)$ of class prototypes a $k$-partition $f(L) = P = (P_1, ..., P_k)$ of objects with classes defined by the minimum-dissimilarity rule:

$$P_h := \{\omega \in \Omega : D(\ell_h, \{\omega\}) \leq D(\ell_m, \{\omega\}), 1 \leq m \leq k\}, \qquad h = 1, \ldots, k (4)$$

where in the case of ties an object $\omega$ is assigned to the prototype (class) with the smallest index.

*The representation function* $g : \mathcal{P}_k \longrightarrow \mathcal{L}_k$ is a function that assigns to each partition $P = (P_1, \ldots, P_k)$ a system $g(P) = L = (\ell_1, \ldots, \ell_k)$ of class representatives in such a way that, for each class $P_h$, $\ell_k$ is the 'most typical element' in $\mathcal{L}_k$ in the sense:

$$\ell_h = argmin_{\ell \in \mathcal{L}} \; D(\ell, P_h) \qquad h = 1, \ldots, k. \tag{5}$$

In order to obtain a well-defined function $g$ we assume that for each subset $P_h \subseteq \Omega$ there exists a unique $\ell \in \mathcal{L}$ that minimizes $D(P_h, \ell)$ in $\mathcal{L}_k$.

Applying iteratively the representation function (4) followed by the assignment function (5) in turn decreases steadily the values $W(P, L)$ of the clustering criterion (1) until a local minimum is attained that depends on the data and, typically, on the initial configuration.

Along the lines of this general approach, different specific clustering approaches have been developed that are distinguished by the choice of the clustering criterion (1), i.e., the dissimilarity function $D$ and the representation space $\mathcal{L}$ which may include, e.g., the centroids (Diday (1972)), the factorial axes (Bock (1974), Ok-Sakun (1975)), a probability law (Schroeder (1976)), etc.

## 3 Symbolic Data

There are numerous practical situations where the information gathered for $n$ individuals $\omega_1, ..., \omega_n$ is too complex to be represented by a classical data table where each individual $\omega_i$ takes exactly one value $x_{ij}$ for each of $p$ variables $j = 1, ..., p$. In particular, there are cases when some variables take more than just one single value for each individual. For example, the time needed for a person (individual) to go to work varies from day to day (between 20 to 40 minutes, say), and the means of transportation that is used may be

different over time (e.g., car or bus or ...). In the first case, the "value" for the variable "time" is an interval (here $[20, 40]$), and in the second case, the variable "transportation mean" is a frequency distribution (for example: car 40%, bus 60%). This type of data is called 'symbolic data'.

Such data arise typically in cases where the investigated objects are not *single* individuals, but *classes* of individuals for which some internal (intra-class) variability must be taken into account. For instance, the variable "gender" of the class "people working in a given enterprise" is described by a frequency distribution: 70% male and 30% female. Another example is provided by situations where there is some inaccuracy or uncertainty in (or after) recording the value of a (classical) variable (e.g., due to measurement errors or 'soft' answers); when describing such variables by a distribution of values, we obtain probabilistic, possibilistic or fuzzy data that constitute another kind of symbolic data.

To process and analyze symbolic data new types of variables have been introduced (Bock and Diday (2000)). A variable is called

– a *set-valued* variable if its "values" are nonempty sets of the underlying domain; more specifically it is called

– a *multi-valued* (categorical) variable if its "values" are finite subsets of the underlying domain (e.g., of an alphabet, of a set of categories,...), and

– an *interval* variable if its "values" are intervals of $\Re^1$. Moreover,

– a *modal* variable is a multi-state variable where, for each object (e.g. a class of individuals) we are given a set of categories and, for each category, a frequency or probability which indicates how frequent, likely, or plausible that category is for this object. In the case where an empirical distribution is given, the variable is called a *histogram variable*.

When looking for the set of all individuals of $\Omega$ (or the set of all objects, classes,...) which share some specific properties the following concept may be useful: A *symbolic object* is a conjunction of conditions on the values taken by the variables, each one of the form $[Y(\omega)\ R\ d]$ where $Y$ is a symbolic variable (observed for all $\omega \in \Omega$), $d$ is a description (i.e., an element of a description space $\mathcal{L}$), and $R$ is a relation between descriptions. As an example consider the symbolic description of a group of people participating in a seminar, defined on the variables age, nationality, sex and staff category:

$$s = [age \in [20, 45]] \wedge [nationality \in \{French, English\}] \wedge [sex \in \{M\}]$$

Inversely, this symbolic object $s$ can be seen as defining the set of all elements from $\Omega$ that realize the stated conditions. This set is called the *extent* of $s$, the approach can be denoted as 'definition by intent'.

In this paper, we consider symbolic data tables which involve (only) interval

variables. More specifically, we consider data tables where each individual (row) $\omega_i$, $i = 1, \ldots, n$ is described by $p$ interval-type variables $Y_j$, $j = 1, \ldots, p$ such that each cell $(i, j)$ of the data matrix contains an interval: Here each

Table 1

| | $Y_1$ | ... | $Y_j$ | ... | $Y_p$ |
|---|---|---|---|---|---|
| $\omega_1$ | $[a_{11}, b_{11}]$ | ... | $[a_{1j}, b_{1j}]$ | ... | $[a_{1p}, b_{1p}]$ |
| ... | ... | | ... | | ... |
| $\omega_i$ | $[a_{i1}, b_{i1}]$ | ... | $[a_{ij}, b_{ij}]$ | ... | $[a_{ip}, b_{ip}]$ |
| ... | ... | | ... | | ... |
| $\omega_n$ | $[a_{n1}, b_{n1}]$ | ... | $[a_{nj}, b_{nj}]$ | ... | $[a_{np}, b_{np}]$ |

$\omega_i \in \Omega$ is described by a vector $x_i = (I_{i1}, \ldots, I_{ip})$ of intervals $I_{ij} = [a_{ij}, b_{ij}], j = 1, \ldots, p$. Notice that we may have $a_{ij} = b_{ij}$ for some $i, j$ (that means that the underlying variable is single-valued for the element $\omega_i$). To each $\omega_i$ we can canonically associate the symbolic object $s_i = [Y_1 \in I_{i1}] \wedge \ldots \wedge [Y_p \in I_{ip}]$.

## 4 Dynamic Clustering for Interval Data

When defining the clustering criterion (1) for interval data, we consider the description space $\mathcal{L} = \mathcal{I}^p$ of $p$-dimensional rectangles in $\Re^p$ and define the dissimilarity $D(\ell, P)$ between a class representative $\ell \in \mathcal{I}^p$ and a class $P \subset \Omega$ by (3) where the dissimilarity $\phi(x_i, \ell)$ between two $p$-dimensional rectangles $\ell = [u, v] = ([u_1, v_1], ..., [u_p, v_p])$ and $x_i = [a_i, b_i] = ([a_{i1}, b_{i1}], ..., [a_{ip}, b_{ip}])$ is defined by

$$\phi(x_i, \ell) := \sum_{j=1}^{p} \left[ |a_{ij} - u_j|^2 + |b_{ij} - v_j|^2 \right]. \tag{6}$$

Obviously, this is the squared Euclidean distance $||(a_i, b_i) - (u, v)||^2$ between the $2p$-dimensional points $(u, v)$ and $(a_i, b_i)$ in $\Re^{2p}$ that characterize the rectangles $\ell$ and $x_i$, respectively. A slight generalization is provided by Minkowski-like distance between rectangles defined by

$$\tilde{\phi}(x_i, \ell) := \sum_{j=1}^{p} \left[ |a_{ij} - u_j|^q + |b_{ij} - v_j|^q \right]^{\frac{1}{q}} \tag{7}$$

with a fixed exponent $q > 1$ (see De Carvalho and Souza (1998)). However, we will only use the definition (6) which results for the choice $q = 2$.

The corresponding representation and the assignment functions are then obtained from (4) and (5) as follows:

The *representation function* $g = g(P_1, ..., P_k)$ assigns to each partition $P = (P_1, ..., P_k)$ of $\Omega$ the system $L = L(P) := (\ell_1, \ldots, \ell_k)$ of $p$-dimensional intervals:

$$\ell_h = ([\bar{a}_{h1}, \bar{b}_{h1}], \ldots, [\bar{a}_{hp}, \bar{b}_{hp}]) \qquad \text{for } h = 1, \ldots, k \tag{8}$$

whose lower and upper boundaries are given by the corresponding averaged boundaries of the data intervals $x_i = [a_i, b_i]$ in the classes $P_h$:

$$\bar{a}_{hj} = \frac{1}{|P_h|} \sum_{\omega_i \in P_h} a_{ij} \qquad \text{and} \qquad \bar{b}_{hj} = \frac{1}{|P_h|} \sum_{\omega_i \in P_h} b_{ij} \tag{9}$$

for $j = 1, \ldots, p$ and $h = 1, \ldots, k$ (see, e.g., Bock (2004)).

The *assignment function* $f = f((\ell_1, \ldots, \ell_k)) := (P_1, \ldots, P_k) =: P$ defined by (4) is given here by the minimum-distance rule

$$P_h = \{\omega_i \in \Omega : \varphi(x_i, \ell_h) \leq \varphi(x_i, \ell_m), 1 \leq m \leq k\} \qquad h = 1, ..., k \tag{10}$$

with the distance measure (6) (in case of ties the cluster with lowest index is chosen).

*Example:* In case of the symbolic data table:

| | Age | Weight [kg] |
|---|---|---|
| $\omega_1$ | $[30, 40]$ | $[48, 55]$ |
| $\omega_2$ | $[10, 20]$ | $[30, 45]$ |
| $\omega_3$ | 20 | $[45, 50]$ |

Table 2: A 3 by 2 interval-type data table

we find, e.g., $\varphi(x_1, x_3) = 534.0721$ and the representative (prototype) rectangle of the class $P_h = \{\omega_1, \omega_3\}$ is $\ell_h = [u, v] = ([25, 30], [46.5, 52.5])$.

Inserting the partial solution (8) of the minimization problem (2) in (1), we see that the clustering problem (2) reduces to minimizing the criterion

$$W(P, L(P)) = \sum_{h=1}^{k} \sum_{\omega_i \in P_h} \sum_{j=1}^{p} \left[ (a_{ij} - \bar{a}_{hj})^2 + (b_{ij} - \bar{b}_{hj})^2 \right] \tag{11}$$

by a suitable choice of the partition $P$. Both will be (approximately) solved by the dynamic clustering algorithm described above.

## 5 Three standardization methods for interval data

It is has been stressed above that clustering results and dissimilarity values are strongly affected when modifying the scales of variables. Typically, some standardization must be performed prior to the clustering process. In the case of interval data we standardize the variables separately, each one in a linear way with the same transformation for both the lower and the upper bound of each interval. In the following we describe three alternative standardization methods.

### *5.1 Standardization using the dispersion of the interval centers*

The first method considers the mean and the dispersion of the interval centers (midpoints) and standardizes such that the resulting transformed midpoints have zero mean and dispersion 1 in each dimension.

Formally, for a given variable $j$, let $I_{ij} = [a_{ij}, b_{ij}]$, $i = 1, \ldots, n$, the intervals to be standardized. The mean value and the dispersion of all interval midpoints $(a_{ij} + b_{ij})/2)$ are given by

$$m_j := \frac{1}{n}\sum_{i=1}^{n}\frac{a_{ij}+b_{ij}}{2} = \frac{1}{2}(\bar{a}_j + \bar{b}_j) \text{ and } \; s_j^2 := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{a_{ij}+b_{ij}}{2} - m_j\right)^2, \quad (12)$$

respectively. With this notation, the data interval $I_{ij}$ is transformed into the interval $I_{ij}^{'} = [a_{ij}^{'}, b_{ij}^{'}]$ with boundaries

$$a_{ij}^{'} := \frac{a_{ij} - m_j}{s_j} \qquad \text{and} \qquad b_{ij}^{'} := \frac{b_{ij} - m_j}{s_j} \quad (13)$$

where automatically $a_{ij}^{'} \leq b_{ij}^{'}$ for all $i, j$. As stated above, the midpoints of the new intervals $I_{ij}^{'}$ are standardized with

$$m_j^{'} := \frac{1}{n}\sum_{i=1}^{n}\frac{a_{ij}^{'}+b_{ij}^{'}}{2} = 0 \quad (14)$$

and

$$s_j^{'2} := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{a_{ij}^{'}+b_{ij}^{'}}{2} - m_j^{'}\right)^2 = \frac{1}{s_j^2}\,\frac{1}{n}\sum_{i=1}^{n}\left(\frac{a_{ij}+b_{ij}}{2} - m_j\right)^2 = 1. \quad (15)$$

*Remark:* Note that the mean value $m_j$ and the dispersion $s_j^2$ verify the usual property that $m_j$ minimizes the mean quadratic deviation $\Phi_j(c) = \frac{1}{n}\sum_{i=1}^{n}[(a_{ij}+b_{ij})/2 - c]^2$ with minimum value $s_j^2 = min_c\ \Phi_j(c) = \Phi(m_j)$.

### *5.2 Standardization using the dispersion of the interval boundaries*

The second standardization method transforms the intervals $I_{ij}$ such that the mean and the (joint) dispersion of the rescaled interval boundaries are 0 and 1, respectively. For a given variable $j$ we define the joint dispersion by

$$\tilde{s}_j^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(a_{ij}-m_j)^2 + (b_{ij}-m_j)^2}{2} \tag{16}$$

and transform, for $i = 1, ..., n$, the intervals $I_{ij} = [a_{ij}, b_{ij}]$ to $I_{ij}^{'} = [a_{ij}^{'}, b_{ij}^{'}]$ with

$$a_{ij}^{'} = \frac{a_{ij}-m_j}{\tilde{s}_j} \quad \text{and} \quad b_{ij}^{'} = \frac{b_{ij}-m_j}{\tilde{s}_j} \tag{17}$$

with $a_{ij}^{'} \leq b_{ij}^{'}$ for all $i, j$. Similarly as before, the new intervals $I_{ij}^{'} = [a_{ij}^{'}, b_{ij}^{'}]$ are standardized according to (14) and

$$\tilde{s}_j^{'2} = \frac{1}{2n}\sum_{i=1}^{n}\left[(a_{ij}^{'}-m_j^{'})^2 + (b_{ij}^{'}-m_j^{'})^2\right] = \frac{1}{2n}\sum_{i=1}^{n}\left[(a_{ij}^{'})^2 + (b_{ij}^{'})^2\right] = 1. \tag{18}$$

*Remark:* Similarly as before, the mean value $m_j$ minimizes the joint quadratic deviation $\tilde{\phi}(c) = [\sum_{i=1}^{n}[(a_{ij}-c)^2 + (b_{ij}-c)^2]/(2n)$ and the joint dispersion is identical to the minimum value: $\tilde{s}_j^2 = \min_c\ \tilde{\Phi}(c) = \tilde{\Phi}(m_j)$ for $j = 1, ..., p$.

### *5.3 Standardization using the global range*

Our third standardization method transforms, for a given variable, the intervals $I_{ij} = [a_{ij}, b_{ij}]$ $(i = 1, ..., n)$ such that the range of the $n$ rescaled intervals is the unit interval $[0, 1]$. Let $Min_j = Min\{a_{1j}, ..., a_{nj}\}$ and $Max_j = Max\{b_{1j}, ..., b_{nj}\}$ the extremal lower and upper boundary values. With this notation, we transform the interval $I_{ij} = [a_{ij}, b_{ij}]$ in the interval $I_{ij}' = [a_{ij}', b_{ij}']$ with boundaries

$$a_{ij}^{'} = \frac{a_{ij}-Min_j}{Max_j - Min_j} \quad \text{and} \quad b_{ij}^{'} = \frac{b_{ij}-Min_j}{Max_j - Min_j} \tag{19}$$

with $a'_{ij} \leq b'_{ij}$ as before. Obviously we have here

$$Min\{a'_{1j}, ..., a_{nj}\} = 0 \quad \text{and} \quad Max\{b'_{1j}, ..., b_{nj}\} = 1$$

as desired.

## 6 Cluster Interpretation

Celeux *et al.* (1989) have introduced a family of indices for cluster interpretation for classical quantitative data. These indices come within the framework of the dynamic clustering algorithm, using the $L_2$ distance, and are based on the notion of inertia. In this paper we present suitable adaptations of these indices to interval data.

### *6.1 Measures based on Inertia*

Let $P = (P_1, \ldots, P_k)$ be the final partition of $\Omega = \{\omega_1, \ldots, \omega_n\}$ in $k$ clusters, $n_h$ the cardinal of cluster $h$, and $\ell_h = (J_h^1, \ldots, J_h^p)$ the representative of cluster $h$, with $J_h^j = [\bar{a}_{hj}, \bar{b}_{hj}]$,where $\bar{a}_{hj} = \dfrac{1}{n_h} \sum\limits_{\omega_i \in P_h} a_{ij}$ and $\bar{b}_{hj} = \dfrac{1}{n_h} \sum\limits_{\omega_i \in P_h} b_{ij}$,
$h = 1, \ldots, k, j = 1, \ldots, p$.

Let $\bar{a}_j = \dfrac{1}{n} \sum\limits_{i=1}^{n} a_{ij} = \dfrac{1}{n} \sum\limits_{h=1}^{k} n_h \bar{a}_{hj}$ and $\bar{b}_j = \dfrac{1}{n} \sum\limits_{i=1}^{n} b_{ij} = \dfrac{1}{n} \sum\limits_{h=1}^{k} n_h \bar{b}_{hj}$,
$j = 1, \ldots, p$. Then the global mean vector is defined as $G = (I_1, \ldots, I_p)$ with $I_j = [\bar{a}_j, \bar{b}_j], j = 1, \ldots, p$.

#### *6.1.1 Global inertia*

The global inertia of the data set is defined as

$$\begin{aligned} T &= \sum_{i=1}^{n} \varphi(x_i, G) = \sum_{i=1}^{n} \sum_{j=1}^{p} [(a_{ij} - \bar{a}_j)^2 + (b_{ij} - \bar{b}_j)^2] = \sum_{j=1}^{p} T_j = \\ &= \sum_{h=1}^{k} T^h = \sum_{j=1}^{p} \sum_{h=1}^{k} T_j^h \end{aligned} \tag{20}$$

with

$$T_j^h = \sum_{\omega_i \in P_h} [(a_{ij} - \bar{a}_j)^2 + (b_{ij} - \bar{b}_j)^2] \tag{21}$$

$$T_j = \sum_{i=1}^{n} [(a_{ij} - \bar{a}_j)^2 + (b_{ij} - \bar{b}_j)^2] = \sum_{h=1}^{k} T_j^h \tag{22}$$

and

$$T^h = \sum_{\omega_i \in P_h} \varphi(x_i, G) = \sum_{j=1}^{p} \sum_{\omega_i \in P_h} [(a_{ij} - \bar{a}_j)^2 + (b_{ij} - \bar{b}_j)^2] = \sum_{j=1}^{p} T_j^h \tag{23}$$

$h = 1, \ldots, k, j = 1, \ldots, p.$

### *6.1.2 Within-class inertia*

The within-class inertia measures the internal dispersion of each cluster; it is defined as :

$$\begin{aligned} W &= \sum_{h=1}^{k} \sum_{\omega_i \in P_h} \varphi(x_i, \ell_h) = \sum_{h=1}^{k} \sum_{\omega_i \in P_h} \sum_{j=1}^{p} [(a_{ij} - \bar{a}_{hj})^2 + (b_{ij} - \bar{b}_{hj})^2] = \\ &= \sum_{j=1}^{p} W_j = \sum_{h=1}^{k} W^h = \sum_{j=1}^{p} \sum_{h=1}^{k} W_j^h \end{aligned} \tag{24}$$

with

$$W_j^h = \sum_{\omega_i \in P_h} [(a_{ij} - \bar{a}_{hj})^2 + (b_{ij} - \bar{b}_{hj})^2] \tag{25}$$

$$W_j = \sum_{h=1}^{k} \sum_{\omega_i \in P_h} [(a_{ij} - \bar{a}_{hj})^2 + (b_{ij} - \bar{b}_{hj})^2] = \sum_{h=1}^{k} W_j^h \tag{26}$$

and

$$W^h = \sum_{\omega_i \in P_h} \varphi(x_i, \ell_h) = \sum_{j=1}^{p} \sum_{\omega_i \in P_h} [(a_{ij} - \bar{a}_{hj})^2 + (b_{ij} - \bar{b}_{hj})^2] = \sum_{j=1}^{p} W_j^h \tag{27}$$

$h = 1, \ldots, k, j = 1, \ldots, p.$

### *6.1.3 Between-class inertia*

The between-class inertia measures the dispersion of the cluster representatives; it is defined as:

$$B = \sum_{h=1}^{k} n_h \varphi(\ell_h, G) = \sum_{h=1}^{k} \sum_{j=1}^{p} n_h [(\bar{a}_{hj} - \bar{a}_j)^2 + (\bar{b}_{hj} - \bar{b}_j)^2] =$$
$$= \sum_{j=1}^{p} B_j = \sum_{h=1}^{k} B^h = \sum_{j=1}^{p} \sum_{h=1}^{k} B_j^h \tag{28}$$

with

$$B_j^h = n_h [(\bar{a}_{hj} - \bar{a}_j)^2 + (\bar{b}_{hj} - \bar{b}_j)^2] \tag{29}$$

$$B_j = \sum_{h=1}^{k} [(\bar{a}_{hj} - \bar{a}_j)^2 + (\bar{b}_{hj} - \bar{b}_j)^2] = \sum_{h=1}^{k} B_j^h \tag{30}$$

and

$$B^h = n_h \sum_{j=1}^{p} [(\bar{a}_{hj} - \bar{a}_j)^2 + (\bar{b}_{hj} - \bar{b}_j)^2] = \sum_{j=1}^{p} B_j^h \tag{31}$$

$h = 1, \ldots, k, j = 1, \ldots, p.$

It is easy to establish that $T = W + B$, $T_j = B_j + W_j$, $T^h = B^h + W^h$ and $T_j^h = B_j^h + W_j^h$ , $h = 1 \ldots, k, j = 1, \ldots, p.$

### *6.2 Interpretation indices*

The indices presented in this section are the suitable adaptations of the indices presented in (Celeux *et al.* (1989)) for the classical case, where the data are vectors of $\Re^p$. All these indices range from 0 to 1.

### *6.2.1 General index*

The proportion of inertia explained by the partition is :

$$R = \frac{B}{T} \tag{32}$$

The algorithm is designed so as to maximize $R$. The greater the value of $R$, the more homogeneous are the clusters, and the better the elements of a cluster $P_h$ are represented by its representative $\ell_h,\ h = 1, \ldots, k$.

### *6.2.2 Variable contribution*

The proportion of inertia of variable $j$ taken into account by the partition is

$$COR(j) = \frac{B_j}{T_j} \tag{33}$$

By comparing the value of $COR(j)$ with the value of the general index $R$, which measures the mean discriminant power of all variables, it may be evaluated whether the discriminant power of variable $j$ is above or below the mean.

The relative contribution of variable $j$ to the between-class inertia is given by

$$CTR(j) = \frac{B_j}{B} \tag{34}$$

A high value of $CTR(j)$ indicates that variable $j$ has an important contribution to the separation of the clusters, it varies a lot from cluster to cluster. In general, $CTR(j)$ varies together with $COR(j)$. An interesting case arises when $COR(j)$ has a low value and $CTR(j)$ is high, meaning that variable $j$ has a low discriminant power although it has an important contribution to inertia (Celeux *et al.* (1989).

### *6.2.3 Cluster description*

The proportion of the global inertia explained by cluster $P_h$ is

$$T(h) = \frac{T^h}{T} \tag{35}$$

The relative contribution of cluster $P_h$ to the between-class inertia is

$$B(h) = \frac{B^h}{B} \tag{36}$$

A high value of $B(h)$ indicates that cluster $P_h$ is very distant from the global center.

The relative contribution of cluster $P_h$ to the within-class inertia is

$$W(h) = \frac{W^h}{W} \tag{37}$$

A high value of $W(h)$ indicates that cluster $P_h$ is not homogeneous.

### *6.2.4 Cluster description by variables*

The proportion of the discriminant power of variable $j$ taken into account by cluster $P_h$ is given by

$$COR(j,h) = \frac{B_j^h}{T_j} \tag{38}$$

A high value of $COR(j,h)$ shows that variable $j$ has an homogeneous behaviour among members of cluster $h$.

The relative contribution of variable $j$ to the between-class inertia explained by cluster $P_h$ is given by

$$CTR(j,h) = \frac{B_j^h}{B^h} \tag{39}$$

The relative contribution of variable $j$ and cluster $P_h$ to the between-class inertia is

$$CE(j,h) = \frac{B_j^h}{B} \tag{40}$$

If $CE(j,h)$ is close to 1, then variable $j$ has a high contribution to the distance between the representative of $P_h$ and the global mean vector, $G$, that is, for variable $j$, cluster $P_h$ is rather eccentric.

*Remark:*

For classical quantitative data, where each interval is reduced to one point, the values of $T, T_j, T^h, T_j^h, W, W_j, W^h, W_j^h, B, B_j, B^h$ and $B_j^h,\ j = 1,\ \ldots,\ p,\ h = 1,\ \ldots,\ k$, are multiplied by a factor of 2. Consequently, the global index, the indices of variable contribution, cluster description and cluster description by variables, have exactly the same values as those computed in the classical case.

*6.3 Cluster description by Symbolic Objects*

At the end of the algorithm, each cluster $P_h$ of the final partition is represented by the symbolic object associated to its representative $\ell_h, h = 1, \ldots, k$:

$$t_h = [Y_1 \in J_h^1] \wedge \ldots \wedge [Y_p \in J_h^p] \tag{41}$$

where $J_h^j = [\bar{a}_{hj}, \bar{b}_{hj}], \quad j = 1, \ldots, p, \ h = 1, \ldots, k.$

Moreover, each cluster may be described by the join (Ichino and Yaguchi (1994)) of the symbolic objects associated to its members, that is

$$z_h = [Y_1 \in I_h^1] \wedge \ldots \wedge [Y_p \in I_h^p] \tag{42}$$

where $I_h^j = [Min_{\omega_i \in P_h} a_{ij}, Max_{\omega_i \in P_h} b_{ij}], \quad j = 1, \ldots, p, \ h = 1, \ldots, k.$

Notice, however, that this description, providing a generalization of the cluster, may be influenced by *outliers*, and two descriptions may overlap even if no member of the corresponding clusters do.

Let us also consider

$$t_h \wedge t_{h'} = [Y_1 \in R_1] \wedge \ldots \wedge [Y_p \in R_p] \tag{43}$$

where $R_j = [Max\{\bar{a}_{hj}, \bar{a}_{h'j}\}, Min\{\bar{b}_{hj}, \bar{b}_{h'j}\}]$ if $Max\{\bar{a}_{hj}, \bar{a}_{h'j}\} \leq Min\{\bar{b}_{hj}, \bar{b}_{h'j}\}$, otherwise, $R_j = \emptyset, \ j = 1, \ldots, p; \ h, \ h^{'} = 1, \ldots, k$. Analogously for $z_h \wedge z_{h'}$.

Two matrices, $A$ and $B$ may then be computed, respectively for representatives and joint descriptions, that allow to evaluate the degree of overlap between cluster descriptions:

$$A_{k \times k} = (\alpha_{h\ h'}), \quad \alpha_{h\ h'} = \frac{\ln \pi(t_h \wedge t_{h'})}{\ln \pi(t_{h'})}, \tag{44}$$

$$B_{k \times k} = (\beta_{h\ h'}), \quad \beta_{h\ h'} = \frac{\ln \pi(z_h \wedge z_{h'})}{\ln \pi(z_{h'})} \tag{45}$$

where $\pi(u) = \prod_{j=1}^{p} \left[1 + (\bar{u}_j - \underline{u}_j)\right]$ if $u = [Y_1 \in [\underline{u}_1, \bar{u}_1]] \wedge \ldots \wedge [Y_p \in [\underline{u}_p, \bar{u}_p]]$ is a symbolic object. We have $0 \leq \alpha_{h\ h'} \leq 1$ and $0 \leq \beta_{h\ h'} \leq 1, \ h, h^{'} = 1, \ldots, k.$

High values of $\alpha_{h\ h'}$ and $\beta_{h\ h'}$ indicate that the corresponding clusters are not well separated. High values in $B$ indicate that the corresponding clusters may

have some overlap in the borders, whereas high values in $A$ result from overlap in the representatives, meaning that the two clusters are not separated.

## 7 Experimental results

The method described above has been applied to simulated and real data, so as to study its performance.

### *7.1 Simulated data*

We simulated two standard quantitative data sets in $\Re^2$. Each data set has 450 points scattered among four clusters of unequal sizes: two clusters sizes 150, one cluster with size 50 and one cluster with size 100. The data points of each cluster in each data-set were drawn according to a bi-variate normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$ represented by:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\ \sigma_1\sigma_2 \\ \rho\ \sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

Data-set 1 shows well-separated clusters (Figure 1).

Fig. 1. Seed Points for data-set 1

The data points of each cluster in this data set were drawn according to the following parameters:

- Cluster 1: $\mu_1 = 5$, $\mu_2 = 250$, $\sigma_1 = 5$, $\sigma_2 = 30$, $\rho = 0.7$;

- Cluster 2: $\mu_1 = 35$, $\mu_2 = 320$, $\sigma_1 = 5$, $\sigma_2 = 30$, $\rho = 0.8$;
- Cluster 3: $\mu_1 = 25$, $\mu_2 = 200$, $\sigma_1 = 5$, $\sigma_2 = 5$, $\rho = -0.7$;
- Cluster 4: $\mu_1 = 5$, $\mu_2 = 400$, $\sigma_1 = 5$, $\sigma_2 = 5$, $\rho = -0.8$.

Data-set 2 shows overlapping clusters (Figure 2):

Fig. 2. Seed Points for data-set 2

The data points of each cluster in this second data set were drawn according to the following parameters:

- Cluster 1: $\mu_1 = 5$, $\mu_2 = 250$, $\sigma_1 = 5$, $\sigma_2 = 30$, $\rho = 0.7$;
- Cluster 2: $\mu_1 = 25$, $\mu_2 = 320$, $\sigma_1 = 5$, $\sigma_2 = 30$, $\rho = 0.8$;
- Cluster 3: $\mu_1 = 25$, $\mu_2 = 250$, $\sigma_1 = 5$, $\sigma_2 = 5$, $\rho = -0.7$;
- Cluster 4: $\mu_1 = 10$, $\mu_2 = 350$, $\sigma_1 = 5$, $\sigma_2 = 5$, $\rho = -0.8$.

Each data point $(x_1, x_2)$ in Figures 1 and 2 is a seed of a vector of intervals (rectangle): $([x_1 - \gamma_1/2, x_1 + \gamma_1/2], [x_2 - \gamma_2/2, x_2 + \gamma_2/2])$. The parameters $\gamma_1, \gamma_2$ are randomly selected from the same predefined interval. The intervals considered in this paper are: $[1, 8]$, $[1, 16]$, $[1, 24]$, $[1, 32]$, and $[1, 40]$ (see Figures 3, 4, 5 and 6).

The evaluation of these clustering methods was performed in the framework of a Monte Carlo experience: 100 replications are considered for each interval data set, as well as for each predefined interval. The mean and the standard deviation of the corrected Rand (CR) index (Hubert and Arabie (1985)) for these 100 replications are computed. In each replication the clustering method is run (until convergence to a stationary value of the adequacy criterion $W$) 50 times and the best result, according to the criterion $W$, is selected.

The CR index assesses the degree of agreement (similarity) between an *a priori* partition (in our case, the partition defined by the seed points) and a partition

Fig. 3. Non-standardized interval data, data-set 1

Fig. 4. Interval data standardized using the dispersion of the interval centers, data-set 1

provided by the clustering algorithm. The CR index was used because it is not sensitive to the number of clusters in the partitions neither to the distributions of the objects in the clusters.

If $U = \{u_1, \ldots, u_r, \ldots, u_R\}$ is a partition obtained by the clustering algorithm, and $V = \{v_1, \ldots, v_c, \ldots, v_C\}$ is a *a priori* partition, the CR index between $U$ and $V$ is defined as:

$$\mathrm{CR} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{R}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}\right] - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}} \tag{46}$$

where $n_{ij}$ denotes the number of objects in clusters $u_i$ and $v_j$; $n_{i.}$ indicates the

Fig. 5. Non-standardized interval data, data-set 2

Fig. 6. Interval data standardized using the dispersion of the interval centers, data-set 2

number of objects in cluster $u_i$; $n_{.j}$ indicates the number of objects in cluster $v_j$; and $n$ is the total number of objects.

The index CR can take values in the interval [-1,1], where the value 1 indicates a perfect agreement between the partitions, whereas values close to 0 (or negative) corresponds to cluster agreements found by chance.

Table 2 shows the values of the mean and the standard deviation of the CR index values for the different method and seed points, for the first data-set (well separated clusters).

Table 3 shows the values of the mean and the standard deviation of the CR index values for the different methods and seed points, for the second data-set (overlapping clusters).

The values in tables 2 and 3 show that standardization, performed by either

Table 2
Mean and Standard Deviation of the CR index values for the different methods and seed points, for the first data-set

| Predefined | No stand. | | Stand. 1 | | Stand. 2 | | Stand. 3 | |
|---|---|---|---|---|---|---|---|---|
| intervals | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| $[1, 8]$ | 0.595 | 0.00123 | 0.956 | 0.00078 | 0.954 | 0.00080 | 0.942 | 0.00745 |
| $[1, 16]$ | 0.595 | 0.00164 | 0.949 | 0.00275 | 0.949 | 0.00346 | 0.926 | 0.0137 |
| $[1, 24]$ | 0.589 | 0.00185 | 0.956 | 0.00080 | 0.954 | 0.00253 | 0.918 | 0.0145 |
| $[1, 32]$ | 0.591 | 0.00166 | 0.946 | 0.00481 | 0.929 | 0.0111 | 0.928 | 0.0113 |
| $[1, 40]$ | 0.597 | 0.00177 | 0.943 | 0.00351 | 0.936 | 0.00728 | 0.927 | 0.00788 |

Table 3
Mean and Standard Deviation of the CR index values for the different methods and seed points, for the second data-set

| Predefined | No stand. | | Stand. 1 | | Stand. 2 | | Stand. 3 | |
|---|---|---|---|---|---|---|---|---|
| intervals | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| $[1, 8]$ | 0.359 | 0.00048 | 0.645 | 0.00887 | 0.639 | 0.01245 | 0.605 | 0.0172 |
| $[1, 16]$ | 0.357 | 0.00049 | 0.651 | 0.00844 | 0.597 | 0.0183 | 0.560 | 0.0193 |
| $[1, 24]$ | 0.358 | 0.00044 | 0.656 | 0.01015 | 0.548 | 0.0198 | 0.559 | 0.0181 |
| $[1, 32]$ | 0.360 | 0.00086 | 0.648 | 0.00956 | 0.523 | 0.0189 | 0.518 | 0.0156 |
| $[1, 40]$ | 0.357 | 0.00049 | 0.638 | 0.0105 | 0.525 | 0.0134 | 0.496 | 0.0136 |

of the three proposed methods, greatly improves the clustering results, as expected. For data-set 1, with well-separated clusters, the method applied to standardized data produced very good results, returning partitions which are very close to the *a priori* partition.

Among the three standardization methods, the standardization using the dispersion of the interval centers produces results that are always slightly better than those produced by the two other methods, while standardization using the global range produced slightly worse results in each case.

### *7.2 The 'Car' example*

The 'car' data-set consists of a set of 33 car models described by 8 interval, 2 categorical multi-valued and one nominal variables (see table 4). In this application, only the 8 interval variables - *Price, Engine Capacity, Top Speed, Acceleration, Step, Length, Width* and *Height* - have been considered for clustering purposes, the nominal variable *Car Category* has been used as a *a priori* classification.

The dynamic clustering method described in Section 4 has been applied to this data-set for $k = 4$ clusters, first without a prior standardization, and then applying each of the three standardization methods introduced in Section 5.

Table 4
'Car' data-set with 8 interval and one nominal variables

| | Price | Engine Capacity | ... | Height | Category |
|---|---|---|---|---|---|
| Alfa 145 | [27806, 33596] | [1370, 1910] | ... | [143, 143] | Utilitarian |
| Alfa 156 | [41593, 62291] | [1598, 2492] | ... | [142, 142] | Berlina |
| ... | ... | ... | ... | ... | ... |
| Porsche 25 | [147704, 246412] | [3387, 3600] | ... | [130, 131] | Sporting |
| Rover 25 | [21492, 33042] | [1119, 1994] | ... | [142, 142] | Utilitarian |
| Passat | [39676, 63455] | [1595, 2496] | ... | [146, 146] | Luxury |

The *a priori* classification, indicated by the suffix attached to the car model denomination, is as follows:

`Utilitarian:`
1-Alfa 145/U 5-Audi A3/U 12-Punto/U 13-Fiesta/U 17-Lancia Y/U
24-Nissan Micra/U 25-Corsa/U 28-Twingo/U 29-Rover 25/U 31-Skoda Fabia/U

`Berlina:`
2-Alfa 156/B 6-Audi A6/B 8-BMW serie 3/B 14-Focus/B
21-Mercedes Classe C/B 26-Vectra/B 30-Rover 75/B 32-Skoda Octavia/B

`Sporting:`
4-Aston Martin/S 11-Ferrari/S 15-Honda NSK/S 16-Lamborghini/S
19-Maserati GT/S 20-Mercedes SL/S 27-Porsche/S

`Luxury:`
3-Alfa 166/L 7-Audi A8/L 9-BMW serie 5/L 10-BMW serie 7/L
18-Lancia K/L 22-Mercedes Classe E/L 23-Mercedes Classe S/L 33-Passat/L

The results of the clustering method are summarized in table 5, where Stand. 1 denotes standardization using the dispersion of the interval centers, Stand. 2 denotes standardization using the dispersion of the interval bounds, and Stand. 3 denotes standardization using the global range.

The clustering method produced the same partition with all three standardization methods, which is rather different than that obtained without standardization. This partition gathers all *Berlina* cars in one Cluster (Cluster 1), all *Sporting* cars in another cluster (Cluster 3), and forms one cluster with most *Utilitarian* cars (Cluster 2) and another with most *Luxury* cars (Cluster 4). It is clear that the partition obtained with standardization of the data is more compatible with the *a priori* categorization than that obtained without standardization.

From table 7 we can see that in the partition obtained without standardization, cluster 4 is the closest to the global mean vector (see Section 6.1), while cluster 3 (three *Sporting* cars) is the farthest. On the other hand, cluster 2 is the least homogeneous of the four clusters, and cluster 1 presents a low value for $W(h)$, due to the low cardinal of this cluster.

Table 5
Clustering Results for the Car Data-Set

| Method | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| No stand. | 22/L 23/L | 1/U 2/B 3/L<br>5/U 6/B 8/B<br>12/U 13/U 14/B<br>17/U 18/L 21/B<br>24/U 25/U 26/B<br>28/U 29/U 30/B<br>31/U 32/B 33/L | 4/S 11/S 16/S | 7/L 9/L 10/L<br>15/S 19/S 20/S<br>27/S |
| Stand. 1 | 1/U 2/B 3/L<br>5/U 8/B 14/B<br>18/L 21/B 26/B<br>30/B 32/B 33/L | 12/U 13/U 17/U<br>24/U 25/U 28/U<br>29/U 31/U | 4/S 11/S 15/S<br>16/S 19/S 20/S<br>27/S | 6/B 7/L 9/L<br>10/L 22/L 23/L |
| Stand. 2 | 1/U 2/B 3/L<br>5/U 8/B 14/B<br>18/L 21/B 26/B<br>30/B 32/B 33/L | 12/U 13/U 17/U<br>24/U 25/U 28/U<br>29/U 31/U | 4/S 11/S 15/S<br>16/S 19/S 20/S<br>27/S | 6/B 7/L 9/L<br>10/L 22/L 23/L |
| Stand. 3 | 1/U 2/B 3/L<br>5/U 8/B 14/B<br>18/L 21/B 26/B<br>30/B 32/B 33/L | 12/U 13/U 17/U<br>24/U 25/U 28/U<br>29/U 31/U | 4/S 11/S 15/S<br>16/S 19/S 20/S<br>27/S | 6/B 7/L 9/L<br>10/L 22/L 23/L |

Table 6
General index for all methods

| Method | R |
|---|---|
| No standardization | 0.916779 |
| Standardization using dispersion of interval centers | 0.776215 |
| Standardization using dispersion of interval bounds | 0.775035 |
| Standardization using global range | 0.778167 |

Table 7
Indices of cluster description for method without standardization

| Cluster | Cardinal | $T(h)$ | $B(h)$ | $W(h)$ |
|---|---|---|---|---|
| 1 | 2 | 0.147393 | 0.158525 | 0.024760 |
| 2 | 21 | 0.285315 | 0.278015 | 0.365734 |
| 3 | 3 | 0.473741 | 0.490851 | 0.285249 |
| 4 | 7 | 0.093552 | 0.072609 | 0.324256 |

The values in table 8 show that in the partition obtained when standardization is performed using the dispersion of the interval centers, cluster 1 is the closest to the global mean, while cluster 3 (all *Sporting* cars) is the farthest and also the least homogeneous of the four clusters.

Tables 9 and 10 show that, when performing standardization using the dispersion of the interval bounds, or using the global range, clusters 2 and 3 are

Table 8
Indices of cluster description for method with standardization using the dispersion of the interval centers

| Cluster | Cardinal | $T(h)$ | $B(h)$ | $W(h)$ |
|---|---|---|---|---|
| 1 | 12 | 0.093634 | 0.054376 | 0.229805 |
| 2 | 8 | 0.308889 | 0.363858 | 0.118225 |
| 3 | 7 | 0.421324 | 0.391660 | 0.524216 |
| 4 | 6 | 0.176152 | 0.190105 | 0.127755 |

Table 9
Indices of cluster description for method with standardization using the dispersion of the interval bounds

| Cluster | Cardinal | $T(h)$ | $B(h)$ | $W(h)$ |
|---|---|---|---|---|
| 1 | 12 | 0.092931 | 0.052087 | 0.233643 |
| 2 | 8 | 0.311504 | 0.366390 | 0.122412 |
| 3 | 7 | 0.414615 | 0.384202 | 0.519389 |
| 4 | 6 | 0.180951 | 0.197320 | 0.124557 |

Table 10
Indices of cluster description for method with standardization using the global range

| Cluster | Cardinal | $T(h)$ | $B(h)$ | $W(h)$ |
|---|---|---|---|---|
| 1 | 12 | 0.094847 | 0.056044 | 0.230967 |
| 2 | 8 | 0.310026 | 0.367230 | 0.109358 |
| 3 | 7 | 0.412902 | 0.380926 | 0.525071 |
| 4 | 6 | 0.182225 | 0.195801 | 0.134603 |

the most eccentric, and cluster 1 is very close to the global center, $G$. Cluster 2 (*Utilitarian* cars) is the most homogeneous and cluster 3 (all *Sporting* cars) the least homogeneous of the four clusters.

Table 11
Indices of partition and cluster description by variables for method without standardization (%)

| | Partition | | Cluster 1 | | Cluster 2 | | Cluster 3 | | Cluster 4 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | COR | CTR | COR | CTR | COR | CTR | COR | CTR | COR | CTR |
| Price | 91.7 | 99.99 | 14.5 | 99.99 | 25.5 | 99.98 | 45.0 | 99.98 | 6.66 | 99.97 |
| Engine Capacity | 79.8 | 0.01 | 9.63 | 0.01 | 23.6 | 0.01 | 37.3 | 0.01 | 9.24 | 0.02 |
| Top Speed | 74.2 | 0.0 | 0.41 | 0.0 | 21.9 | 0.0 | 37.6 | 0.0 | 14.3 | 0.0 |
| Acceleration | 62.7 | 0.0 | 1.81 | 0.0 | 20.7 | 0.0 | 24.4 | 0.0 | 15.7 | 0.0 |
| Step | 28.0 | 0.0 | 19.4 | 0.0 | 5.73 | 0.0 | 0.07 | 0.0 | 2.78 | 0.0 |
| Length | 29.9 | 0.0 | 10.8 | 0.0 | 9.75 | 0.0 | 2.80 | 0.0 | 6.57 | 0.0 |
| Width | 61.6 | 0.0 | 2.99 | 0.0 | 17.9 | 0.0 | 33.9 | 0.0 | 6.78 | 0.0 |
| Height | 62.3 | 0.0 | 1.85 | 0.0 | 11.7 | 0.0 | 43.1 | 0.0 | 5.59 | 0.0 |

Comparing the values of $COR(j)$ with the value of $R$ (see table 6) , for the partition obtained without standardization, we may conclude that variable

*Price*'s discriminant power is equivalent to the mean value, while all other variables have a discriminant power below the mean. The values in table 11 also allow to conclude that variable *Price* has the most important role in the formation of the clusters, being almost totally responsible for the inter-class inertia. We'll see that this effect will disappear with standardization.

Table 12
Indices of partition and cluster description by variables with standardization using the dispersion of the interval centers (%)

| | Partition | | Cluster 1 | | Cluster 2 | | Cluster 3 | | Cluster 4 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | COR | CTR | COR | CTR | COR | CTR | COR | CTR | COR | CTR |
| Price | 75.4 | 13.1 | 11.1 | 35.5 | 16.5 | 7.91 | 37.9 | 16.9 | 9.90 | 9.08 |
| Engine Capacity | 75.5 | 13.1 | 7.68 | 24.5 | 21.8 | 10.4 | 31.7 | 14.1 | 14.3 | 13.0 |
| Top Speed | 87.1 | 13.6 | 2.63 | 7.57 | 33.4 | 14.3 | 49.0 | 19.5 | 2.02 | 1.66 |
| Acceleration | 83.5 | 14.1 | 1.31 | 4.08 | 39.1 | 18.1 | 38.2 | 16.5 | 4.81 | 4.27 |
| Step | 75.2 | 11.6 | 0.66 | 1.87 | 28.3 | 12.0 | 2.20 | 0.87 | 44.0 | 35.8 |
| Length | 80.3 | 12.3 | 1.23 | 3.47 | 49.1 | 20.6 | 1.12 | 0.44 | 28.9 | 23.3 |
| Width | 69.1 | 10.7 | 0.94 | 2.67 | 35.8 | 15.2 | 22.0 | 8.65 | 10.4 | 8.42 |
| Height | 74.9 | 11.5 | 7.19 | 20.4 | 3.40 | 1.44 | 58.7 | 23.1 | 5.55 | 4.50 |

Table 13
Indices of partition and cluster description by variables with standardization using the dispersion of the interval bounds (%)

| | Partition | | Cluster 1 | | Cluster 2 | | Cluster 3 | | Cluster 4 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | COR | CTR | COR | CTR | COR | CTR | COR | CTR | COR | CTR |
| Price | 75.4 | 11.9 | 11.1 | 35.5 | 16.5 | 7.91 | 37.9 | 16.9 | 9.90 | 9.08 |
| Engine Capacity | 75.5 | 11.3 | 7.68 | 24.5 | 21.8 | 10.4 | 31.7 | 14.1 | 14.3 | 13.0 |
| Top Speed | 87.1 | 14.0 | 2.63 | 7.57 | 33.4 | 14.3 | 49.0 | 19.5 | 2.02 | 1.66 |
| Acceleration | 83.5 | 12.2 | 1.31 | 4.08 | 39.1 | 18.1 | 38.2 | 16.5 | 4.81 | 4.27 |
| Step | 75.2 | 12.7 | 0.66 | 1.87 | 28.3 | 12.0 | 2.20 | 0.87 | 44.0 | 35.8 |
| Length | 80.3 | 13.6 | 1.23 | 3.47 | 49.1 | 20.6 | 1.12 | 0.44 | 28.9 | 23.3 |
| Width | 69.1 | 11.7 | 0.94 | 2.67 | 35.8 | 15.2 | 22.0 | 8.65 | 10.4 | 8.42 |
| Height | 74.9 | 12.6 | 7.19 | 20.4 | 3.40 | 1.44 | 58.7 | 23.1 | 5.55 | 4.50 |

Comparing the values of $COR(j)$ with the values of $R$ (see table 6), for the partitions obtained using either of the standardization methods (see tables 12, 13 and 14) we may conclude that the discriminant power of variables *Top Speed*, *Acceleration* and *Length* is higher than the mean value, while all other variables have a discriminant power slightly below the mean. The values in tables 12, 13 and 14 also allow to conclude that variables *Price*, *Engine Capacity* and *Height* are the most important in the formation of Cluster 1, variable *Length* is the most important in the formation of Cluster 2, variable *Top Speed* is the most important in the formation of Cluster 3 and variables

Table 14
Indices of partition and cluster description by variables with standardization using the global range (%)

| | Partition | | Cluster 1 | | Cluster 2 | | Cluster 3 | | Cluster 4 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | COR | CTR | COR | CTR | COR | CTR | COR | CTR | COR | CTR |
| Price | 75.4 | 13.9 | 11.1 | 36.3 | 16.5 | 8.25 | 37.9 | 18.3 | 9.90 | 9.28 |
| Engine Capacity | 75.5 | 16.0 | 7.68 | 29.1 | 21.8 | 12.6 | 31.7 | 17.7 | 14.3 | 15.5 |
| Top Speed | 87.1 | 14.5 | 2.63 | 7.84 | 33.4 | 15.2 | 49.0 | 21.5 | 2.02 | 1.72 |
| Acceleration | 83.5 | 12.6 | 1.31 | 3.53 | 39.1 | 16.0 | 38.2 | 15.1 | 4.81 | 3.70 |
| Step | 75.2 | 11.6 | 0.66 | 1.81 | 28.3 | 11.9 | 2.20 | 0.89 | 44.0 | 34.7 |
| Length | 80.3 | 13.6 | 1.23 | 3.72 | 49.1 | 22.5 | 1.12 | 0.50 | 28.9 | 24.9 |
| Width | 69.1 | 8.80 | 0.94 | 2.14 | 35.8 | 12.4 | 22.0 | 7.35 | 10.4 | 6.75 |
| Height | 74.9 | 9.09 | 7.19 | 15.6 | 3.40 | 1.12 | 58.7 | 18.7 | 5.55 | 3.44 |

*Step* and *Length* are the most important in the formation of Cluster 4. With standardization, irrespective to the technique used, variable *Price* loses the importance it had when no standardization was applied, the responsibility for the cluster formation being now distributed by the whole set of variables.

Table 15
Values of $CE(j,h)$ for each cluster and variable, for methods without standardization and with standardization using the dispersion of the interval centers (Stand. 1) (%)

| | No Stand. | | | | Stand. 1 | | | |
|---|---|---|---|---|---|---|---|---|
| | Cl. 1 | Cl. 2 | Cl. 3 | Cl. 4 | Cl. 1 | Cl. 2 | Cl. 3 | Cl. 4 |
| Price | 15.9 | 27.8 | 49.1 | 7.26 | 1.93 | 2.88 | 6.61 | 1.73 |
| Engine Capacity | 0.0 | 0.0 | 0.0 | 0.0 | 1.33 | 3.77 | 5.50 | 2.48 |
| Top Speed | 0.0 | 0.0 | 0.0 | 0.0 | 0.41 | 5.21 | 7.66 | 0.32 |
| Acceleration | 0.0 | 0.0 | 0.0 | 0.0 | 0.22 | 6.60 | 6.46 | 0.81 |
| Step | 0.0 | 0.0 | 0.0 | 0.0 | 0.10 | 4.37 | 0.34 | 6.80 |
| Length | 0.0 | 0.0 | 0.0 | 0.0 | 0.19 | 7.51 | 0.17 | 4.43 |
| Width | 0.0 | 0.0 | 0.0 | 0.0 | 0.14 | 5.52 | 3.39 | 1.60 |
| Height | 0.0 | 0.0 | 0.0 | 0.0 | 1.11 | 0.52 | 9.04 | 0.86 |

Tables 15 and 16 show that without standardization, only variable *Price* contributes to the separation of the four clusters, while when either of the standardization techniques is applied, no variable presents a high value of $CE(j,h)$ for any of the clusters: separation between cluster representatives is due the whole set of variables. Nevertheless, it may be said that, without standardization, cluster 3, with three *Sporting* cars, is the most eccentric. For the other three methods (i.e. with standardization), cluster 2 (*Utilitarian* cars) is a little more eccentric as respects variables *Length* and *Width*; cluster 3, comprehending only *Sporting* cars, as respects variables *Top Speed* and *Height*; and cluster 4, with only *Luxury* cars, as respects variable *Step*.

Table 16
Values of $CE(j,h)$ for each cluster and variable, for methods with standardization using the dispersion of the interval bounds (Stand. 2) and using the global range (Stand. 3) (%)

| | Stand. 2 | | | | Stand. 3 | | | |
|---|---|---|---|---|---|---|---|---|
| | Cl. 1 | Cl. 2 | Cl. 3 | Cl. 4 | Cl. 1 | Cl. 2 | Cl. 3 | Cl. 4 |
| Price | 1.75 | 2.61 | 5.91 | 1.57 | 2.03 | 3.03 | 6.96 | 1.82 |
| Engine Capacity | 1.15 | 3.26 | 4.76 | 2.14 | 1.63 | 4.62 | 6.75 | 3.03 |
| Top Speed | 0.42 | 5.38 | 7.90 | 0.33 | 0.44 | 5.57 | 8.18 | 0.34 |
| Acceleration | 0.19 | 5.72 | 5.60 | 0.70 | 0.20 | 5.88 | 5.76 | 0.72 |
| Step | 0.11 | 4.78 | 0.37 | 7.43 | 0.10 | 4.37 | 0.34 | 6.80 |
| Length | 0.21 | 8.28 | 0.19 | 4.88 | 0.21 | 8.28 | 0.19 | 4.88 |
| Width | 0.16 | 6.04 | 3.71 | 1.75 | 0.12 | 4.56 | 2.80 | 1.32 |
| Height | 1.21 | 0.57 | 9.91 | 0.94 | 0.87 | 0.41 | 7.13 | 0.67 |

The symbolic objects associated to the representatives of the clusters, for the partition obtained without standardization, are as follows:

$t_1 = [Price \in [98722.5, 391873.5]] \wedge [Engine\ Capacity \in [1506.6, 2207.2]] \wedge \ldots \wedge [Height \in [144.0, 144.0]]$

$t_2 = [Price \in [35144.9, 55052.3]] \wedge [Engine\ Capacity \in [2598.5, 5612.5]] \wedge \ldots \wedge [Height \in [143.0, 143.0]]$

$t_3 = [Price \in [304597.3, 424897.3]] \wedge [Engine\ Capacity \in [5171.0, 5800.3]] \wedge \ldots \wedge [Height \in [121.7, 124.3]]$

$t_4 = [Price \in [134254.1, 218665.0]] \wedge [Engine\ Capacity \in [2873.6, 4278.6]] \wedge \ldots \wedge [Height \in [135.7, 135.9]]$

The generalizing descriptions for the partition obtained without standardization are as follows:

$z_1 = [Price \in [69243, 394342]] \wedge [Engine\ Capacity \in [1998, 5786]] \wedge \ldots \wedge [Height \in [144, 144]]$

$z_2 = [Price \in [16992, 140265]] \wedge [Engine\ Capacity \in [973, 4172]] \wedge \ldots \wedge [Height \in [132, 148]]$

$z_3 = [Price \in [240292, 460000]] \wedge [Engine\ Capacity \in [3586, 5992]] \wedge \ldots \wedge [Height \in [111, 132]]$

$z_4 = [Price \in [70292, 276792]] \wedge [Engine\ Capacity \in [2171, 5987]] \wedge \ldots \wedge [Height \in [129, 144]]$

Matrices $A$ and $B$ for the partition obtained without standardization are as follows:

$$A = \begin{pmatrix} 1.0 & 0.0 & 0.96 & 0.99 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.85 & 0.0 & 1.0 & 0.45 \\ 0.89 & 0.0 & 1.0 & 1.0 \end{pmatrix} \quad B = \begin{pmatrix} 1.0 & 0.85 & 0.91 & 0.88 \\ 0.91 & 1.0 & 0.59 & 0.89 \\ 0.92 & 0.48 & 1.0 & 0.74 \\ 0.97 & 0.88 & 0.83 & 1.0 \end{pmatrix}$$

The values in matrix $A$ show that Cluster 2, which gathers the least expensive cars, is well separated from the other clusters; Cluster 1 overlaps with both

Cluster 3 and Cluster 4, and these latter also present some overlap. Matrix $B$ shows that the generalizing descriptions of the clusters intersect in a great extent.

The symbolic objects associated to the representatives of the clusters, for the partition obtained with standardization, are as follows:

$t_1 = [Price \in [42984.3, 65595.7]] \wedge [Engine\ \ Capacity \in [1707.9, 2424.3]] \wedge \ldots \wedge [Height \in [143.2, 143.2]]$

$t_2 = [Price \in [19251.9, 28585.6]] \wedge [Engine\ \ Capacity \in [1170.3, 1636.0]] \wedge \ldots \wedge [Height \in [142.6, 142.6]]$

$t_3 = [Price \in [222076.9, 308335.2]] \wedge [Engine\ \ Capacity \in [3984.7, 4769.1]] \wedge \ldots \wedge [Height \in [126.3, 127.6]]$

$t_4 = [Price \in [94115.7, 261835.5]] \wedge [Engine\ \ Capacity \in [2452.2, 4894.0]] \wedge \ldots \wedge [Height \in [144.0, 144.0]]$

The generalizing descriptions for the partition obtained with standardization are as follows:

$z_1 = [Price \in [27419, 115248]] \wedge [Engine\ \ Capacity \in [1370, 3199]] \wedge \ldots \wedge [Height \in [142, 146]]$

$z_2 = [Price \in [16992, 33042]] \wedge [Engine\ \ Capacity \in [973, 1994]] \wedge \ldots \wedge [Height \in [132, 148]]$

$z_3 = [Price \in [132800, 460000]] \wedge [Engine\ \ Capacity \in [2799, 5992]] \wedge \ldots \wedge [Height \in [111, 132]]$

$z_4 = [Price \in [68216, 394342]] \wedge [Engine\ \ Capacity \in [1781, 5786]] \wedge \ldots \wedge [Height \in [143, 145]]$

Matrices $A$ and $B$ for the partition obtained with standardization are as follows:

$$A = \begin{pmatrix} 1.0 & 0.0 & 0.0 & 0.26 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.87 \\ 0.29 & 0.0 & 0.96 & 1.0 \end{pmatrix} \quad B = \begin{pmatrix} 1.0 & 0.77 & 0.78 & 0.87 \\ 0.72 & 1.0 & 0.39 & 0.65 \\ 0.83 & 0.43 & 1.0 & 0.95 \\ 0.98 & 0.60 & 0.89 & 1.0 \end{pmatrix}$$

The values in matrix $A$ show that Cluster 2, which comprehends only *Utilitarian* cars, is well separated from the other clusters; Cluster 3 and Cluster 4, with *Sporting* and *Luxury* cars, respectively, overlap in great extent; Cluster 1, which also contains some *Luxury* cars, seems to be close to cluster 4. Again, matrix $B$ shows that the generalizing descriptions of the clusters intersect in a great extent; Cluster 2 an Cluster 3, with *Utilitarian* and *Sporting* cars, respectively, are the better separated clusters.

## 8 Concluding remarks

In this paper, a method for clustering interval data, based on the dynamic clustering methodology, is proposed. The method is designed to consider clas-

sical quantitative and interval variables together, allowing for the presence of degenerate intervals. The distance used to assign objects to clusters is a Minkowski-like distance, of $L_2$ type. Each cluster is represented by a vector of intervals, whose bounds are, for each variable, respectively, the mean of the lower bounds and the mean of the upper bounds, computed for cluster. The final partition corresponds to a local optimum of a criterion, which evaluates the fit between the clusters and their representatives.

Special attention is given to the issue of standardization, and three standardization techniques, adapted to interval data, are introduced.

The problem of interpreting and evaluating the obtained partition has then been addressed. A family of indices, based on the notion of inertia are presented, which constitute suitable adaptations to interval data of indices introduced by (Celeux *et al.* (1989)).

The simulation study carried out showed that standardization greatly improves the performance of the clustering method, and that, among the three proposed standardization techniques, the standardization using the dispersion of the interval centers produces results that are always slightly better than those produced by the two other methods. The application to the real dataset stressed the need for standardization prior to the clustering process and allowed to explore the potentialities of the proposed family of interpretation indices.

Further developments concern the extension of the method to generalized symbolic data.

## References


Bock, H.-H. 1974. Automatische Klassifikation. Vandenhoeck & Ruprecht, Goettingen, chapter 17.

Bock, H.-H. 2002. Clustering algorithms and Kohonen maps for symbolic data. In: Journal of the Japanese Society of Computational Statistics, 15, 1-13.

Bock, H.-H. 2004. Visualizing symbolic data by Kohonen maps. In: M. Noirhomme, E. Diday (eds.): Symbolic data analysis and the SODAS-ASSO software. Dekker, New York, 2004 (in press).

Bock, H.-H. and Diday, E. (eds.) 2000. Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data. Springer, Heidelberg.

Celeux, G.; Diday, E. ; Govaert, G. ; Lechevallier, Y. ; Ralambondrainy, H. 1989. Classification Automatique des Données. Bordas, Paris.

Chavent, M. and Lechevallier, Y. 2002. Dynamical Clustering Algorithm of Interval Data: Optimization of an Adequacy Criterion Based on Hausdorff

Distance. In: Sokolowsky and H.H. Bock Eds., Classification, Clustering and Data Analysis. Springer, Heidelberg, 53-59.

De Carvalho, F. A. T. and Souza, R. M. C. R. 1998. New metrics for Constrained Boolean Symbolic Objects. In: Studies and Reserach: Proceedings of the Conference on Knowledge Extraction and Symbolic Data Analysis (KESDA'98), Office for Official Publications of the European Communities, Luxemburg, 175-187.

Diday, E. 1972. Nouveaux Concepts et Nouvelles Méthodes en Classification Automatique, Thèse d'Etat, Univ. Paris VI.

Diday, E. and Simon, J. J. 1976. Clustering Analysis. In: K. S. Fu Eds., Digital Pattern Recognition. Springer, Heidelberg, 47-94.

Diday, E. and Brito, M. P. 1989. Symbolic Cluster Analysis. In: O. Opitz Eds., Conceptual and Numerical Analysis of Data. Springer-Verlag, Heidelberg, 45-84.

El-Sonbaty, Y. and Ismail, M. A. 1998. Fuzzy Clustering for Symbolic Data. IEEE Transactions on Fuzzy Systems 6, 195-204.

Gordon, A. D. 1999. Classification 2nd edition, Chapman & Hall, Boca Raton (Florida).

Gordon, A. D. 2000. An Iteractive Relocation Algorithm for Classifying Symbolic Data In: W. Gaul *et al* Eds., Data Analysis: Scientific Modeling and Practical Application. W. Gaul, Springer-Verlag, Berlin, 17-23.

Hubert, L. and Arabie. P. 1985. Comparing Partitions. Journal of Classification 2, 193-218.

Ichino, M. and Yaguchi, H. 1994. Generalized Minkowski metrics for mixed feature type data analysis. IEEE Transactions on Systems, Man and Cybernetics, 24, (4), 698-708.

Jain, A.K., Murty, M.N. and Flynn, P.J. 1999. Data Clustering: A Review. ACM Computing Surveys, 31, (3), 264-323.

Ok-Sakun, Y. 1975. Analyse Factorielle Typologique et Lissage Typologique, Thèse de 3ème cycle, Univ. Paris VI.

Ralambondrainy, H. 1995. A Conceptual version of the K-means algorithm. Pattern Recognition Letters 16, 1147-1157.

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

Schroeder, A. 1976. Analyse d'un mélange de distributions de probabilité de même type. R.S.A. Vol. 24, 1.

Verde, R., De Carvalho, F. A. T. and Lechevallier, Y. 2001. A Dynamical Clustering Algorithm for symbolic data. Tutorial on Symbolic Data Analisys, GfKl Conference, Munich.