# Dynamic Cluster Methods for Interval Data based on Mahalanobis Distances

Renata M.C.R. de Souza[1], Francisco de A.T. de Carvalho[1], Camilo P. Tenório[1] and Yves Lechevallier[2]

[1] Centro de Informatica - CIn / UFPE, Av. Prof. Luiz Freire, s/n - Cidade Universitaria, CEP: 50740-540 - Recife - PE - Brasil, Email: {fatc,rmcrs,cpt}@cin.ufpe.br

[2] INRIA - Rocquencourt, Domaine de Voluceau - Rocquencourt - B. P. 105 78153 Le Chesnay Cedex - France, Email: Yves.Lechevallier@inria.fr

**Summary.** Dynamic cluster methods for interval data are presented. Two methods are considered: the first method furnish a partition of the input data and a corresponding prototype (a vector of intervals) for each class by optimizing an adequacy criterion which is based on Mahalanobis distances between vectors of intervals. The second is an adaptive version of the first method. Experimental results with artificial interval data sets show the usefulness of these methods. Otherwise, the adaptive method outperforms the non-adaptive one concerning the quality of the clusters which are furnished by the algorithms.

## 1 Introduction

Cluster analysis have been widely used in numerous fields including pattern recognition, data mining and image processing. Their aim is to group data into clusters such that objects within a cluster have high degree of similarity whereas objects belonging to different clusters have high degree of dissimilarity.

The dynamic cluster algorithm (Diday and Simon (1976)) is a partitional clustering method whose aim is to obtain both a single partition of the input data and the identification of a suitable representation or prototype (means, axes, probability laws, groups of elements, etc.) for each cluster by locally minimizing an adequacy criterion which measures the fitting between the clusters and their representation. The k-means algorithm with class prototype updated after all objects have been considered for relocation, is a particular case of dynamic clustering with adequacy function equal to squared error criterion such that class prototypes equal to clusters centers of gravity (Jain et al. (1999)).

In the adaptive version of the dynamic cluster method (Diday and Govaert (1977)), at each iteration there is a different measure to the comparison of each cluster with its own representation. The advantage of these adaptive distances is that the clustering algorithm is able to recognize clusters of different shapes and sizes.

Often, objects to be clustered are represented as a vector of quantitative features. However, the recording of interval data has become a common practice in real world applications and nowadays this kind of data is widely used to describe objects. Symbolic Data Analysis (SDA) is a new area related to multivariate analysis and pattern recognition, which has provided suitable data analysis methods for managing objects described as a vector of intervals (Bock and Diday (2000)).

Concerning partitioning clustering methods, SDA has provided suitable tools. Verde et al. (2001) introduced a dynamic cluster algorithm for interval data considering context dependent proximity functions. Chavent and Lechevallier (2002) proposed a dynamic cluster algorithm for interval data using an adequacy criterion based on Hausdorff distance. Souza and De Carvalho (2004) presented dynamic cluster algorithms for interval data based on adaptive and non-adaptive city-block distances.

The main contribution of this paper is to introduce two dynamic cluster methods for interval data. The first method furnishes a partition of the input data and a corresponding prototype (a vector of intervals) for each class by optimizing an adequacy criterion which is based on Mahalanobis distances between vectors of intervals (section 2). The second is an adaptive version of the first method (section 3). In both methods, the prototype of each cluster is represented by a vector of intervals, where the bounds of each interval are respectively, for a fixed variable, the average of the set of lower bounds and the average of the set of upper bounds of the intervals of the objects belonging to the cluster for the same variable. In order to show the usefulness of these methods, several artificial interval data sets ranging from different degree of difficulty to be clustered were considered. The evaluation of the clustering results is based on an external validity index in the framework of a Monte Carlo experience (section 4). Finally, in section 5 are given the conclusions.

## 2 A dynamic cluster with non-adaptive Mahalanobis distance for interval data

Let $E = \{s_1, \ldots, s_n\}$ be a set of $n$ symbolic objects described by $p$ interval variables. Each object $s_i$ $(i = 1, \ldots, n)$ is represented as a vector of intervals $\mathbf{x}_i = ([a_i^1, b_i^1], \ldots, [a_i^p, b_i^p])^T$. Let $P$ be a partition of $E$ into $K$ clusters $C_1, \ldots, C_K$, where each cluster $C_k$ $(k = 1, \ldots, K)$ has a prototype $L_k$ that is also represented as a vector of intervals $\mathbf{y}_k = ([\alpha_k^1, \beta_k^1], \ldots, [\alpha_k^p, \beta_k^p])^T$.

According to the standard dynamic cluster algorithm, our method look for a partition $P = (C_1, \ldots, C_K)$ of a set of objects into $K$ clusters and its corresponding set of prototypes $L = (L_1, \ldots, L_K)$ by locally minimizing an adequacy criterion usually defined in the following way:

$$W_1(P,L)=\sum_{k=1}^{K}\Delta_k^1(L_k)=\sum_{k=1}^{K}\sum_{i\,\in\,C_k}\delta(\mathbf{x}_i,\mathbf{y}_k) \tag{1}$$

where $\delta(\mathbf{x}_i,\mathbf{y}_k)$ is a distance measure between an object $s_i\in C_k$ and the class prototype $L_k$ of $C_k$.

Let $\mathbf{x}_{iL}=(a_i^1,\dots,a_i^p)^T$ and $\mathbf{x}_{iU}=(b_i^1,\dots,b_i^p)^T$ be two vectors, respectively, of the lower and upper bounds of the intervals describing $\mathbf{x}_i$. Consider also $\mathbf{y}_{kL}=(\alpha_k^1,\dots,\alpha_k^p)^T$ and $\mathbf{y}_{kU}=(\beta_k^1,\dots,\beta_k^p)^T$ be two vectors, respectively, of the lower and upper bounds of the intervals describing $\mathbf{y}_k$.

We define the distance between the two vectors of intervals $\mathbf{x}_i$ and $\mathbf{y}_k$ as:

$$\delta(\mathbf{x}_i,\mathbf{y}_k)=d(\mathbf{x}_{iL},\mathbf{y}_{kL})+d(\mathbf{x}_{iU},\mathbf{y}_{kU}) \tag{2}$$

where

$$d(\mathbf{x}_{iL},\mathbf{y}_{kL})=(\mathbf{x}_{iL}-\mathbf{y}_{kL})^T\mathbf{M}_L(\mathbf{x}_{iL}-\mathbf{y}_{kL}) \tag{3}$$

is the Mahalanobis distance between the two vectors $\mathbf{x}_{iL}$ and $\mathbf{y}_{kL}$ and,

$$d(\mathbf{x}_{iU},\mathbf{y}_{kU})=(\mathbf{x}_{iU}-\mathbf{y}_{kU})^T\mathbf{M}_U(\mathbf{x}_{iU}-\mathbf{y}_{kU}) \tag{4}$$

is the Mahalanobis distance between the two vectors $\mathbf{x}_{iU}$ and $\mathbf{y}_{kU}$.

The matrices $\mathbf{M}_L$ and $\mathbf{M}_U$ are defined, respectively, as:

(i) $\mathbf{M}_L=(\,\det(\mathbf{Q}_{poolL}))^{1/p}\ \mathbf{Q}_{poolL}^{-1}$, where $\mathbf{Q}_{poolL}$ is the pooled covariance matrix with $\det(\mathbf{Q}_{poolL})\neq 0$, i.e.,

$$\mathbf{Q}_{poolL}=\frac{(n_1-1)\mathbf{S}_{1L}+\ldots+(n_K-1)\mathbf{S}_{KL}}{n_1+\ldots+n_K-K} \tag{5}$$

In equation (5), $\mathbf{S}_{kL}$ is the covariance matrix of the set of vectors $\{\mathbf{x}_{iL}/i\in C_k\}$ and $n_k$ is the cardinal of $C_k$ $(k=1,\dots,K)$.

(ii) $\mathbf{M}_U=(\,\det(\mathbf{Q}_{poolU}))^{1/p}\ \mathbf{Q}_{poolU}^{-1}$, where $\mathbf{Q}_{poolU}$ is the pooled covariance matrix with $\det(\mathbf{Q}_{poolU})\neq 0$, i.e.,

$$\mathbf{Q}_{poolU}=\frac{(n_1-1)\mathbf{S}_{1U}+\ldots+(n_k-1)\mathbf{S}_{KU}}{n_1+\ldots+n_K-K} \tag{6}$$

In equation (6), $\mathbf{S}_{kU}$ is the covariance matrix of the set of vectors $\{\mathbf{x}_{iU}/s_i\in C_k\}$ and $n_k$ is again the cardinal of $C_k$ $(k=1,\dots,K)$.

### 2.1 The optimization problem

In this method the optimization problem is stated as follows: find the vector of intervals $\mathbf{y}_k = ([\alpha_k^1, \beta_k^1], \ldots, [\alpha_k^p, \beta_k^p])$ which locally minimizes the following adequacy criterion:

$$\Delta_k^1(L_k) = \sum_{i \in C_k} (\mathbf{x}_{iL} - \mathbf{y}_{kL})^T \mathbf{M}_L (\mathbf{x}_{iL} - \mathbf{y}_{kL}) + \sum_{i \in C_k} (\mathbf{x}_{iU} - \mathbf{y}_{kU})^T \mathbf{M}_U (\mathbf{x}_{iU} - \mathbf{y}_{kU}) \quad (7)$$

The problem now becomes to find the two vectors $\mathbf{y}_{kL}$ and $\mathbf{y}_{kU}$ minimizing the criterion $\Delta_1^k(L_k)$. According to Govaert (1975), the solution for $\mathbf{y}_{kL}$ and $\mathbf{y}_{kU}$ are obtained from the Huygens theorem. They are, respectively, the mean vector of the sets $\{\mathbf{x}_{iL}/s_i \in C_k\}$ and $\{\mathbf{x}_{iU}/s_i \in C_k\}$.

Therefore, $\mathbf{y}_k$ is a vector of intervals whose bounds are, for each variable j, respectively, the average of the set of lower bounds and the average of the set of upper bounds of the intervals of the objects belonging to the cluster $C_k$.

### 2.2 The algorithm

The dynamic cluster algorithm with non-adaptive Mahalanobis distance has the following steps:

```
1. Initialization. Randomly choose a partition $\{C_1 \ldots, C_K\}$ of $E$.
2. Representation step.
   For $k = 1$ to $K$ compute the vector $\mathbf{y}_k = ([\alpha_k^1, \beta_k^1], \ldots, [\alpha_k^p, \beta_k^p])$
   where $\alpha_k^j$ is the average of $\{a_i^j/s_i \in C_k\}$ and $\beta_k^j$ is the average of $\{b_i^j/s_i \in C_k\}$, $j = 1, \ldots, p$.
3. Allocation step.
   $test \leftarrow 0$

   for $i = 1$ to $n$ do
       define the cluster $C_{k*}$ such that
           $k* = arg\ \min_{k=1,\ldots,K}\ (\mathbf{x}_{iL} - \mathbf{y}_{kL})^T \mathbf{M}_L (\mathbf{x}_{iL} - \mathbf{y}_{kL}) +$
                                  $(\mathbf{x}_{iU} - \mathbf{y}_{kU})^T \mathbf{M}_U (\mathbf{x}_{iU} - \mathbf{y}_{kU})$
       if $i \in C_k$ and $k* \neq k$
           $test \leftarrow 1$
           $C_{k*} \leftarrow C_{k*} \cup \{s_i\}$
           $C_k \leftarrow C_k \setminus \{s_i\}$

4. Stopping criterion.
   If $test = 0$ then STOP, else go to (2).
```

## 3 Dynamical cluster with adaptive Mahalanobis distance for interval data

The dynamic cluster algorithm with adaptive distances (Diday and Govaert (1977)) has also a representation and an allocation step but there is a different distance associated to each cluster. The algorithm looks for a partition in $K$ clusters, its corresponding $K$ prototypes and $K$ different distances associated with the clusters by locally minimizing an adequacy criterion which is usually stated as:

$$W_2(P,L) = \sum_{k=1}^{K} \Delta_k^2(L_k,\delta_k) = \sum_{k=1}^{K} \sum_{i \in C_k} \delta_k(\mathbf{x}_i,\mathbf{y}_k) \tag{8}$$

where $\delta_k(\mathbf{x}_i,\mathbf{y}_k)$ is an adaptive dissimilarity measure between an object $s_i \in C_k$ and the class prototype $L_k$ of $C_k$.

According to the intra-class structure of the cluster $C_k$, we consider here an adaptive Mahalanobis distance between an object $s_i$ and a prototype $L_k$, which is defined as:

$$\begin{aligned}\delta_k(\mathbf{x}_i,\mathbf{y}_k) = {} & (\mathbf{x}_{iL} - \mathbf{y}_{kL})^T \mathbf{M}_{kL} (\mathbf{x}_{iL} - \mathbf{y}_{kL}) + \\ & (\mathbf{x}_{iU} - \mathbf{y}_{kU})^T \mathbf{M}_{kU} (\mathbf{x}_{iU} - \mathbf{y}_{kU})\end{aligned} \tag{9}$$

where $\mathbf{M}_{kL}$ and $\mathbf{M}_{kU.}$ are matrices associated to the cluster $C_k$, both of determinant equal to 1.

### 3.1 The optimization problem

The optimization problem has two stages:

a) The class $C_k$ and the matrices $\mathbf{M}_{kL}$ and $\mathbf{M}_{kU}$ $(k = 1, \ldots, K)$ are fixed. We look for the prototype $L_k$ of the class $C_k$ which locally minimizes

$$\begin{aligned}\Delta_k^2(L_k,\delta_k) = {} & \sum_{i \in C_k} (\mathbf{x}_{iL} - \mathbf{y}_{kL})^T \mathbf{M}_{kL} (\mathbf{x}_{iL} - \mathbf{y}_{kL}) + \\ & \sum_{i \in C_k} (\mathbf{x}_{iU} - \mathbf{y}_{kU})^T \mathbf{M}_{kU} (\mathbf{x}_{iU} - \mathbf{y}_{kU})\end{aligned} \tag{10}$$

As we know from subsection 2.1, the solutions for $\alpha_{kL}^j$ and $\beta_{kU}^j$ are, respectively, the average of $\{a_i^j, s_i \in C_k\}$, the lower bounds of the intervals $[a_i^j, b_i^j]$, $s_i \in C_k$, and the average of $\{b_i^j, s_i \in C_k\}$, the upper bounds of the intervals $[a_i^j, b_i^j]$, $s_i \in C_k$.

b) The class $C_k$ and the prototypes $L_k$ $(k = 1, \ldots, K)$ are fixed.
   We look for the distance $\delta_k$ of the class $C_k$ which locally minimizes the criterion $\Delta_k^2$ with $\det(\mathbf{M}_{kL}) = 1$ and $\det(\mathbf{M}_{kU}) = 1$.
   According to Diday and Govaert (1977), the solutions are: $\mathbf{M}_{kL} =$ (det $\mathbf{Q}_{kL})^{1/p}$ $\mathbf{Q}_{kL}^{-1}$ where $\mathbf{Q}_{kL}$ is the covariance matrix of the lower bounds of the intervals belonging to the class $C_k$ with $\det(\mathbf{Q}_{kL}) \neq 0$ and $\mathbf{M}_{kU} =$ (det $\mathbf{Q}_{kU})^{1/p}$ $\mathbf{Q}_{kU}^{-1}$ where $\mathbf{Q}_{kU}$ is the covariance matrix of the upper bounds of the intervals belonging to the class $C_k$ with $\det(\mathbf{Q}_{kU}) \neq 0$.

### 3.2 The algorithm

The initialization, the allocation step and the stopping criterion are nearly the same in the adaptive and non-adaptive dynamic cluster algorithm. The main difference between these algorithms occurs in the representation step when it is computed for each class $k$, $(k = 1, \ldots, K)$ the matrices $\mathbf{M}_{kL} = (\det(\mathbf{Q}_{kL}))^{1/p}$ $\mathbf{Q}_{kL}^{-1}$ and $\mathbf{M}_{kU} = (\det(\mathbf{Q}_{kU}))^{1/p}$ $\mathbf{Q}_{kU}^{-1}$.

*Remark.* If a single number is considered as an interval with equal lower and upper bounds, the results furnished by these symbolic-oriented methods are identical to those furnished by the standard numerical ones when usual data (vector of single quantitative values) are used. Indeed, the clusters and the respective prototypes are identical.

## 4 Experimental results

To show the usefulness of these methods, experiments with two artificial interval data sets, of different degrees of clustering difficulty (clusters of different shapes and sizes, linearly non-separable clusters, etc), are considered. The experiments have three stages: generation of usual and interval data (stages 1 and 2), and evaluation of the clustering results in the framework of a Monte Carlo experience.

### 4.1 Usual data sets

Initially, we considered two standard quantitative data sets in $\Re^2$. Each data set has 450 points scattered among four clusters of unequal sizes and shapes: two clusters with ellipsis shapes and sizes 150 and two clusters with spherical shapes of sizes 50 and 100. The data points of each cluster in each data set were drawn according to a bi-variate normal distribution with correlated components.

Data set 1, showing well-separated clusters, is generated according to the following parameters:

a) Class 1: $\mu_1 = 28$, $\mu_2 = 22$, $\sigma_1^2 = 100$, $\sigma_{12} = 21$, $\sigma_2^2 = 9$ and $\rho_{12} = 0.7$;
b) Class 2: $\mu_1 = 65$, $\mu_2 = 30$, $\sigma_1^2 = 9$, $\sigma_{12} = 28.8$, $\sigma_2^2 = 144$ and $\rho_{12} = 0.8$;

c) Class 3: $\mu_1 = 45$, $\mu_2 = 42$, $\sigma_1^2 = 9$, $\sigma_{12} = 6.3$, $\sigma_2^2 = 9$ and $\rho_{12} = 0.7$;
d) Class 4: $\mu_1 = 38$, $\mu_2 = -1$, $\sigma_1^2 = 25$, $\sigma_{12} = 20$, $\sigma_2^2 = 25$ and $\rho_{12} = 0.8$;

Data set 2, showing overlapping clusters, is generated according to the following parameters:

a) Class 1: $\mu_1 = 45$, $\mu_2 = 22$, $\sigma_1^2 = 100$, $\sigma_{12} = 21$, $\sigma_2^2 = 9$ and $\rho_{12} = 0.7$;
b) Class 2: $\mu_1 = 65$, $\mu_2 = 30$, $\sigma_1^2 = 9$, $\sigma_{12} = 28.8$, $\sigma_2^2 = 144$ and $\rho_{12} = 0.8$;
c) Class 3: $\mu_1 = 57$, $\mu_2 = 38$, $\sigma_1^2 = 9$, $\sigma_{12} = 6.3$, $\sigma_2^2 = 9$ and $\rho_{12} = 0.7$;
d) Class 4: $\mu_1 = 42$, $\mu_2 = 12$, $\sigma_1^2 = 25$, $\sigma_{12} = 20$, $\sigma_2^2 = 25$ and $\rho_\rho 12 = 0.8$ ;

### 4.2 Interval data sets

Each data point $(z_1, z_2)$ of the data set 1 and 2 is a seed of a vector of intervals (rectangle): $([z_1 - \gamma_1/2, z_1 + \gamma_1/2], [z_2 - \gamma_2/2, z_2 + \gamma_2/2])$. These 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 $[1, 40]$. Figure 1 shows interval data set 1 with well separated clusters and Figure 2 shows interval data set 2 with overlapping clusters.

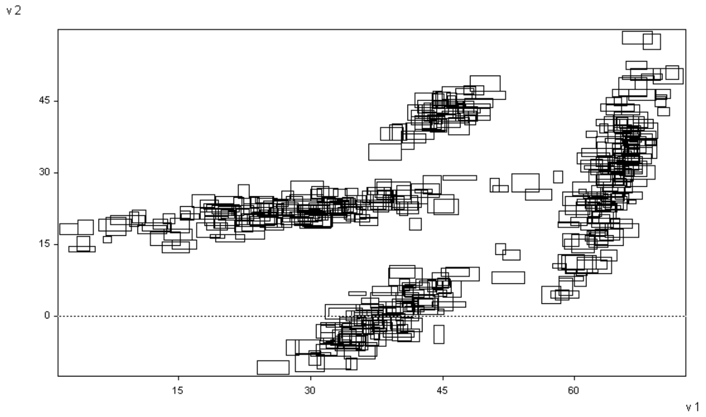

**Fig. 1.** Interval data set 1 showing well-separated classes

### 4.3 The Monte Carlo Experience

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. In each replication a clustering method is run (until the convergence to a stationary value of the adequacy criterion $W_1$ or $W_2$) 50 times and the best result, according to the criterion $W_1$ or $W_2$, is selected.

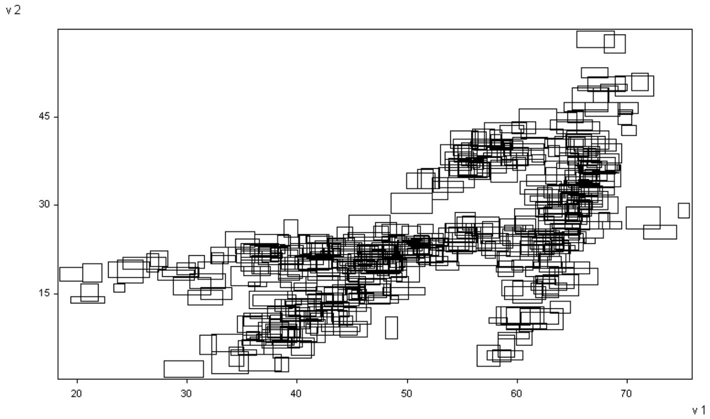

**Fig. 2.** Interval data set 2 showing overlapping classes

*Remark.* As in the standard Mahalanobis (adaptive and non-adaptive) distance methods for dynamic cluster, these methods have sometimes a problem with the inversion of matrices. When this occurs, the actual version of these algorithms stops the current iteration and re-starts a new one. The stopped iteration is not take into account among the 50 which should be run.

The average of the corrected Rand (CR) index (Hubert and Arabie (1985)) among these 100 replications is calculated. 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 furnished by the clustering algorithm.

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

$$\mathrm{CR} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}{\frac{1}{2}[\sum_{i=1}^{R}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}] - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}} \tag{11}$$

where $n_{ij}$ represents the number of objects that are in clusters $u_i$ and $v_i$; $n_{i.}$ indicates the 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.

CR can take values in the interval [-1,1], where the value 1 indicates a perfect agreement between the partitions, whereas values near 0 (or negatives) correspond to cluster agreements found by chance (Milligan (1996)).

Table 1 shows the values of the average CR index according to the different methods and interval data sets. This table also shows suitable (null and alternative) hypothesis and the observed values of statistics following a Student's t distribution with 99 degrees of freedom.

**Table 1.** Comparison between the clustering methods

| Range of values | Interval Data Set 1 | | | Interval Data Set 2 | | |
|---|---|---|---|---|---|---|
| of $\gamma_i$ $i = 1, 2$ | Non-Adaptive Method | Adaptive Method | $H_0 : \mu_1 \leq \mu$ $H_a : \mu_1 > \mu$ | Non-Adaptive Method | Adaptive Method | $H_0 : \mu_1 \leq \mu$ $H_a : \mu_1 > \mu$ |
| $\gamma_i \in [1, 8]$ | 0.778 | 0.996 | 80.742 | 0.409 | 0.755 | 13.266 |
| $\gamma_i \in [1, 16]$ | 0.784 | 0.986 | 82.182 | 0.358 | 0.688 | 22.609 |
| $\gamma_i \in [1, 24]$ | 0.789 | 0.963 | 61.464 | 0.352 | 0.572 | 20.488 |
| $\gamma_i \in [1, 32]$ | 0.802 | 0.937 | 39.181 | 0.349 | 0.435 | 18.204 |
| $\gamma_i \in [1, 40]$ | 0.805 | 0.923 | 29.084 | 0.341 | 0.386 | 9.2851 |

As the interval data set used to calculate the CR index by each method in each replication is exactly the same, the comparison between the proposed clustering methods is achieved by the paired Student's t-test at a significance level of 5%. In these tests, $\mu_1$ and $\mu$ are, respectively, the average of the CR index for adaptive and non-adaptive methods.

From the results in Table 1, it can be seen that the average CR indices for the adaptive method are greater than those for the non-adaptive method in all situations. In addition, the statistic tests support the hypothesis that the average performance (measured by the CR index) of the adaptive method is superior to the non-adaptive method.

## 5 Conclusions

In this paper, dynamic cluster methods for interval data are presented. Two methods are considered: the first method furnish a partition of the input data and a corresponding prototype (a vector of intervals) for each class by optimizing an adequacy criterion which is based on Mahalanobis distances between vectors of intervals. The second is an adaptive version of the first method. In both methods the prototype of each class is represented by a vector of intervals, where the bounds of these intervals for a variable are, respectively, the average of the set of lower bounds and the average of the set of upper bounds of the intervals of the objects belonging to the class for the same variable. The convergence of these algorithms and the decrease of their partitioning criterions at each iteration is due to the optimization of their adequacy criterions at each representation step. The accuracy of the results furnished by these clustering methods were assessed by the corrected Rand index considering artificial interval data sets ranging from different degrees of clustering difficulties in the framework of a Monte Carlo experience. Concerning the average CR index, the method with adaptive distance clearly outperforms the method with non-adaptive distance.

*Acknowledgments*: The authors would like to thank CNPq (Brazilian Agency) for its financial support.

## References

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

Chavent, M. and Lechevallier, Y. (2002). Dynamical Clustering Algorithm of Interval Data: Optimization of an Adequacy Criterion Based on Hausdorff Distance. In *Classification, Clustering and Data Analysis* (S. et al, ed.), 53–59. Springer, Heidelberg.

Diday, E. and Govaert, G. (1977). Classification Automatique avec Distances Adaptatives. *R.A.I.R.O. Informatique Computer Science* **11** 329–349.

Diday, E. and Simon, J. J. (1976). Clustering Analysis. In *Digital Pattern Recognition* (K. S. Fu, ed.), 47–94.

Govaert, G. (1975). *Classification automatique et distances adaptatives.* Ph.D. dissertation, hèse de 3ème cycle, Mathématique appliquée, Université Paris VI.

Hubert, L. and Arabie, P. (1985). Comparing Partitions. *Journal of Classification* **2** 193–218.

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

Milligan, G. W. (1996). Clustering Validation: results and implications for applied analysis. In *Clustering and Classification*, 341–375. Word Scientific, Singapore.

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** 353–365.

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