Partitioning fuzzy c-means clustering algorithms for interval-valued data based on city-block distances

Francisco de A. T. de Carvalho and Gibson B. N. Barbosa
Centro de Informatica (CIn)
Universidade Federal de Pernambuco (UFPE)
Av. Jornalista Anibal Fernandes, s/n - Cidade Universitária
CEP: 50740-560, Recife-PE, Brazil
Email: {fatc,gbnb}@cin.ufpe.br

Julio T. Pimentel
Departamento de Engenharia Mecanica
Universidade Federal de Pernambuco (UFPE)
Av. Jornalista Anibal Fernandes, s/n - Cidade Universitária
CEP: 50740-560, Recife-PE, Brazil
Email: julio.pimentel@gmail.comm

***Abstract*—This paper presents partitioning fuzzy c-means clustering algorithms for interval-valued data based on city-block distaces. These fuzzy c-means clustering algorithms give a fuzzy partition and a prototype for each fuzzy cluster by optimizing an adequacy criterion based on suitable adaptive and non-adaptive city-block distances between vectors of intervals. The adaptive city-block distances change at each algorithm iteration and are different from one fuzzy cluster to another. Experiments with real interval-valued data sets show the usefulness of these fuzzy clustering algorithms.**

Index Terms*—*fuzzy c-means; interval-valued data; city-block distances;

I. Introduction

This paper aims to present partitioning fuzzy clustering algorithms in order to cluster objects described by intervalvalued variables. Interval-valued variables are needed, for example, when an object represents a group of individuals and the variables used to describe it need to assume a value which express the variability inherent to the description of a group. Interval-valued data arise in practical situations such as recording monthly interval temperatures at meteorological stations, daily interval stock prices, etc. Another source of interval-valued data is the aggregation of huge databases into a reduced number of groups, the properties of which are described by interval-valued variables. Therefore, tools for interval-valued data analysis [1], [2] are very much required.

Clustering is an unsupervised learning method widely used in various areas such that data mining, information retrieval, computer vision, bioinformatics, etc. Clustering algorithms aims to organize a set of objects into clusters such that items within a given cluster have a high degree of similarity, while items belonging to different clusters have a high degree of dissimilarity. The most popular clustering techniques are hierarchical and partitioning [3], [4], [5]: hierarchical methods yield a complete hierarchy, i.e., a nested sequence of partitions of the input data, whereas partitioning methods seek to obtain a single partition of the input data into a fixed number of clusters, usually by optimizing an objective function.

Partitioning clustering is divided into hard and fuzzy methods. In hard clustering methods, each object of the data set must be assigned to precisely one cluster. Fuzzy clustering, however, furnishes a fuzzy partition based on the idea of the partial membership of each pattern in a given cluster. Concerning objects described by real-valued variables, Dunn [6] presented the first fuzzy clustering algorithms based on an adequacy criterion defined by Euclidean distances. Bezdek [7] further generalized this method. Diday and Govaert [8] introduced one of the first approaches to using adaptive distances in partitioning quantitative data. Gustafson and Kessel [9] introduced the first adaptive fuzzy clustering algorithm, based on a quadratic distance defined by a full fuzzy covariance matrix estimated locally for each cluster. De Carvalho et al [10] introduced fuzzy K-means clustering algorithms based on adaptive quadratic distances, either defined by full as well as diagonal fuzzy covariance matrices estimated globally or defined by diagonal fuzzy covariance matrices estimated locally for each cluster.

Concerning fuzzy clustering of interval-valued data, El-Sonbaty and Ismail [11] presented a fuzzy K-means algorithm for clustering data on the basis of different types of symbolic variables. Yang et al [12] presented fuzzy clustering algorithms for mixed features of symbolic and fuzzy data. In these fuzzy clustering algorithms, the degree of membership is associated to the values of the features in the clusters for the cluster centers rather than being associated to the patterns in each cluster, as is the standard approaches. De Carvalho [13] presented fuzzy c-means clustering algorithms based on Euclidean distances and De Carvalho and Tenorio [14] introduced fuzzy c-means clustering algorithms based on adaptive quadratic distances.

Theoretical studies indicate that city-block based models are more robust than those based on Euclidean distances. For this purpose, Jajuga [15] introduced fuzzy c-means clusterings algorithms based on non-adaptive city-block distances between vectors of quantitative values. This paper aims to give fuzzy c-means clustering algorithms for interval-valued data based on adaptive and non-adaptive city-block distances.

This paper is organized as follows. Section II presents the fuzzy c-means clustering algorithms based on (adaptive and non-adaptive) City-Block distances. Section III presents experiments with synthetic and real interval-valued data sets in order to show the usefulness of these fuzzy clustering algorithms. Finally, Section IV gives concluding remarks.

978-0-7695-5092-3/13 $26.00 © 2013 IEEE
DOI 10.1109/BRACIS.2013.27

II. Fuzzy c-means clustering algorithm for interval-valued data with city-block distances

This section presents fuzzy c-means clustering algorithms for interval-valued data based on adaptive ($AIFCM-L1$) and non-adaptive ($IFCM-L1$) city-block distances. These algorithms are suitable extensions of the standard fuzzy cmeans clustering algorithm based on city-block distances [15].

Let $E = \{e_1, \ldots, e_n\}$ be a set of n objects indexed by k and described by p interval variables indexed by j. An *interval variable* [1] X is a correspondence defined from E in $\Re$ such that for each $e_k \in E$, $X(e_k) = [a, b] \in \Im$, where $\Im$ is the set of closed intervals defined from $\Re$. Each object e_k is represented as a vector of intervals $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})$, where $x_{kj} = [a_{kj}, b_{kj}] \in \Im = \{[a, b] : a, b \in \Re, a \leq b\}$.

A basic assumption of the presented fuzzy c-means clustering algorithms is that a prototype $\mathbf{g}_i$ of fuzzy cluster $P_i (i = 1, \ldots, c)$ is also represented as a vector of intervals $\mathbf{g}_i = (g_{i1}, \ldots, g_{ip})$, where $g_{ij} = [\alpha_{ij}, \beta_{ij}] \in \Im (j = 1, \ldots, p)$.

The $IFCM - L1$ is an iterative two-steps algorithm (representation and allocation steps) that aims to furnish a fuzzy partition of a set of items in c clusters $P_1, \ldots, P_c$, represented by the matrix of membership degrees $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$ with $\mathbf{U}_k = (u_{k1}, \ldots, u_{kc}) (k = 1, \ldots, n)$, and their corresponding prototypes $\mathbf{g}_1, \ldots, \mathbf{g}_c$ such that a criterion J measuring the fitting between the clusters and their representatives (prototypes) is locally minimized. This criterion is based on a non-adaptive city-block distance between vectors of intervals and is defined as:

$$\begin{aligned} J &= \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m d(\mathbf{x}_k, \mathbf{g}_i) \qquad (1) \\ &= \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} [|a_{kj} - \alpha_{ij}| + |b_{kj} - \beta_{ij}|] \end{aligned}$$

where $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})$ is a vector of intervals describing the $k - th$ item, $\mathbf{g}_i = (g_{i1}, \ldots, g_{ip})$ is a vector of intervals describing the prototype of fuzzy cluster P_i, u_{ik} is the membership degree of item e_k in fuzzy cluster P_i, $m \in]1, +\infty[$ is a parameter that controls the fuzziness of membership for each pattern k, and

$$d(\mathbf{x}_k, \mathbf{g}_i) = \sum_{j=1}^{p} [|a_{kj} - \alpha_{ij}| + |b_{kj} - \beta_{ij}|] \qquad (2)$$

is a non-adaptive city-block distance between item $\mathbf{x}_k$ and prototype $\mathbf{g}_i$.

The $AIFCM - L1$ is an iterative three-steps algorithm (representation, weighting and allocation steps) that looks for a fuzzy partition of a set of items in c fuzzy clusters $P_1, \ldots, P_c$, represented by the matrix of membership degrees $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$ with $\mathbf{u}_k = (u_{k1}, \ldots, u_{kc}) (k = 1, \ldots, n)$, the corresponding c prototypes $\mathbf{g}_1, \ldots, \mathbf{g}_c$ and adaptive city-block distances between vectors of intervals that are different for each cluster, such that a criterion J measuring the fitting between the fuzzy clusters and their representatives (prototypes) is locally minimized. This criterion J is based on an adaptive city-block distance for each fuzzy cluster and is defined as:

$$\begin{aligned} J &= \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m d_{\boldsymbol{\lambda}_i}(\mathbf{x}_k, \mathbf{g}_i) \qquad (3) \\ &= \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} \lambda_{ij} [|a_{kj} - \alpha_{ij}| + |b_{kj} - \beta_{ij}|] \end{aligned}$$

where $\mathbf{x}_k, \mathbf{g}_i, u_{ik}$ and m are defined as before and

$$d_{\boldsymbol{\lambda}_i}(\mathbf{x}_k, \mathbf{g}_i) = \sum_{j=1}^{p} \lambda_{ij} [|a_{kj} - \alpha_{ij}| + |b_{kj} - \beta_{ij}|] \qquad (4)$$

is an adaptive city-block distance between item xk and prototype gi defined for each fuzzy cluster and parameterized by the vectors of weights $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{ip}) (i = 1, \ldots, c)$, which change at each iteration.

During the representation step of $IFCM - L1$, the vectors of membership degrees $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are kept fixed. The adequacy criterion J is minimized with respect to the prototypes. During the representation step of $AIFCM - L1$, the vectors of weights $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_c$ and the vectors of membership degrees $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are kept fixed. The adequacy criterion J is also minimized with respect to the prototypes. For both algorithms, the problem is then to find for $i = 1, \ldots, c$ the prototype $\mathbf{g}_i$ that minimizes

$$\begin{aligned} \Delta(\mathbf{g}_i) &= \sum_{k=1}^{n} (u_{ik})^m d(\mathbf{x}_k, \mathbf{g}_i) \\ &= \sum_{j=1}^{p} \sum_{k=1}^{n} (u_{ik})^m [|a_{kj} - \alpha_{ij}| + |b_{kj} - \beta_{ij}|] \end{aligned}$$

The criterion Δ being aditive, the problem becomes to find for $j = 1, \ldots, p$, the interval $g_{ij} = [\alpha_{ij}, \beta_{ij}]$ which minimizes

$$\sum_{i=1}^{n} (u_{ik})^m [|a_{kj} - \alpha_{ij}| + |b_{kj} - \beta_{ij}|]$$

This yields two minimization problems: find $\alpha \in \Re$ and $\beta_{ij} \in \Re$ that minimizes, respectively,

$$\sum_{k=1}^{n} |(u_{ik})^m a_{kj} - \alpha_{ij} (u_{ik})^m| \quad \text{and} \quad \sum_{k=1}^{n} |(u_{ik})^m b_{kj} - \beta_{ij} (u_{ik})^m|$$

Each of these last two problems brings to the minimization of $\sum_{k=1}^{n} |y_k - a z_k|$, where $y_k = (u_{ik})^m a_{kj}$ (or $y_k = (u_{ik})^m b_{kj}$), $z_k = (u_{ik})^m$ and $a = \alpha_{ij}$ (or $a = \beta_{ij}$) . An algorithmic solution of this problem is known and to solve it the following algorithm can be used [15]:

1) Determining $b_k = y_k / z_k (k = 1, \ldots, n)$;
2) Rearrange the z_k's according to ascending order of b_k's and get $\tilde{z}_1, \ldots, \tilde{z}_n$;
3) Minimize $\sum_{l=1}^{r} |\tilde{z}_l| - \sum_{l=r+1}^{n} |\tilde{z}_l|$ regarding r;
4) If the minimum is negative, take $a = b_r$. If the minimum is positive, take $a = b_{r+1}$. Finally, if the minimum is equal to zero, take $a = (b_r + b_{r+1})/2$.

During the weighting step of $AIFCM-L1$, the prototypes $\mathbf{g}_1,\ldots,\mathbf{g}_c$ and the vectors of membership degrees $\mathbf{u}_1,\ldots,\mathbf{u}_n$ un are kept fixed. The adequacy criterion J is minimized with respect to the vectors of weights.

The vector of weights $\boldsymbol{\lambda}_i (i=1,\ldots,c)$, which minimizes the clustering criterion J under $\lambda_{ij}>0$ and $\prod_{j=1}^{p}\lambda_{ij}=1$, has its components calculated (using the multiplier technique of Lagrange) as follows:

$$\lambda_{ij}=\frac{\left\{\prod_{h=1}^{p}\left[\sum_{k=1}^{n}(u_{ik})^m[|a_{kh}-\alpha_{ih}|+|b_{kh}-\beta_{ih}|]\right]\right\}^{\frac{1}{p}}}{\sum_{k=1}^{n}(u_{ik})^m[|a_{kj}-\alpha_{ij}|+|b_{kj}-\beta_{ij}|]} \tag{5}$$

During the allocation step of $IFCM-L1$, the prototypes $\mathbf{g}_1,\ldots,\mathbf{g}_c$ are kept fixed. The adequacy criterion J is minimized with respect to the vectors of membership degrees. The membership degree $u_{ik}(k=1,\ldots,n)$ of each item e_k in each fuzzy cluster $P_i(i=1,\ldots,c)$, minimizing the clustering criterion J under $u_{ik}\geq 0$ and $\sum_{i=1}^{c}u_{ik}=1$, is computed (using the multiplier technique of Lagrange) according to the following expression:

$$\begin{aligned} u_{ik} &= \left[\sum_{h=1}^{c}\left(\frac{d(\mathbf{x}_k,\mathbf{g}_i)}{d(\mathbf{x}_k,\mathbf{g}_h)}\right)^{\frac{1}{m-1}}\right]^{-1} \\ &= \left[\sum_{h=1}^{c}\left(\frac{\sum_{j=1}^{p}[|a_{kj}-\alpha_{ij}|+|b_{kj}-\beta_{ij}|]}{\sum_{j=1}^{p}[|a_{kj}-\alpha_{hj}|+|b_{kj}-\beta_{hj}|]}\right)^{\frac{1}{m-1}}\right]^{-1} \end{aligned} \tag{6}$$

During the allocation step of $AIFCM-L1$, the prototypes $\mathbf{g}_1,\ldots,\mathbf{g}_c$ and the vectors of weights $\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_c$ are kept fixed. The adequacy criterion J is minimized with respect to the vectors of membership degrees.

The membership degree $u_{ik}(i=1,\ldots,n)$ of each item e_k in each fuzzy cluster $P_i(i=1,\ldots,c)$, minimizing the clustering criterion J under $u_{ik}\geq 0$ and $\sum_{i=1}^{c}u_{ik}=1$, is computed (using the multiplier technique of Lagrange) according to the following expression:

$$\begin{aligned} u_{ik} &= \left[\sum_{h=1}^{c}\left(\frac{d_{\boldsymbol{\lambda}_i}(\mathbf{x}_k,\mathbf{g}_i)}{d_{\boldsymbol{\lambda}_h}(\mathbf{x}_k,\mathbf{g}_h)}\right)^{\frac{1}{m-1}}\right]^{-1} \\ &= \left[\sum_{h=1}^{c}\left(\frac{\sum_{j=1}^{p}\lambda_{ij}[|a_{kj}-\alpha_{ij}|+|b_{kj}-\beta_{ij}|]}{\sum_{j=1}^{p}\lambda_{hj}[|a_{kj}-\alpha_{hj}|+|b_{kj}-\beta_{hj}|]}\right)^{\frac{1}{m-1}}\right]^{-1} \end{aligned} \tag{7}$$

A. Algorithm

The fuzzy c-means clustering algorithm with adaptive and non-adaptive city-block distances for interval-valued data can be summarized as follows:

1) *Initialization*.
Fix the number c of clusters; Fix 1 <m <1;
Fix T (the maximum number of iterations), fix $\epsilon \ll 1$;
Set $t \leftarrow 0$;
Randomly select c distinct prototypes $\mathbf{g}_i^{(0)} \in E_{ik}=(1,\ldots,c)$;
For $AIFCM-L1$ algorithm, set $\boldsymbol{\lambda}_i^{(0)}=(\lambda_{i1}^{(0)},\ldots,\lambda_{ip}^{(0)})=(1,\ldots,1)(i=1,\ldots,c)$;
For $IFCM-L1$ algorithm, compute the components of the vectors of membership degrees: $\mathbf{u}_k^{(0)}=(u_{k1}^{(0)},\ldots,u_{kc}^{(0)})(k=1,\ldots,n)$ according to equation (6).
For $AIFCM-L1$ algorithm, compute the components of the vectors of membership degrees:$\mathbf{u}_k^{(0)}=(u_{k1}^{(0)},\ldots,u_{kc}^{(0)})(k=1,\ldots,n)$ according to equation (7).
For $IFCM-L1$ algorithm, compute $J^{(0)}$ according to equation (1).
For $AIFCM-L1$ algorithm, compute $J^{(0)}$ according to equation (3).

2) *Representation step: computation of the best prototypes*.
Compute the boundaries of the intervals $g_{ij}^{(t)}=[\alpha_{ij}^{(t)},\beta_{ij}^{(t)}](j=1,\ldots,p)$, components of the prototypes $\mathbf{g}_i^{(t)}=(g_{i1}^{(t)},\ldots,g_{ip}^{(t)})(k=1,\ldots,c)$, from

$$\alpha_{ij}^{(t)}=\underset{\alpha_{ij}\in\Re}{\operatorname{argmin}}\sum_{k=1}^{n}|(u_{ik}^{(t-1)})^m a_{kj}-\alpha_{ij}(u_{ik}^{(t-1)})^m|$$

and

$$\beta_{ij}^{(t)}=\underset{\beta_{ij}\in\Re}{\operatorname{argmin}}\sum_{k=1}^{n}|(u_{ik}^{(t-1)})^m b_{kj}-\beta_{ij}(u_{ik}^{(t-1)})^m|$$

according to the algorithm described in section II.

3) *Weighting step: computation of the best weights*
Skip this step for $IFCM-L1$ algorithm.
For $AIFCM-L1$ algorithm, compute the components of the vectors of weights $\boldsymbol{\lambda}_i^{(t)}=(\lambda_{i1}^{(t)},\ldots,\lambda_{ip}^{(t)})(i=1,\ldots,c)$ according to equation (5).

4) *Allocation step: definition of the best fuzzy partition*
For $IFCM-L1$ algorithm, compute the components of the vectors of membership degrees: $\mathbf{u}_k^{(0)}=(u_{k1}^{(0)},\ldots,u_{kc}^{(0)})(k=1,\ldots,n)$ according to equation (6).
For $AIFCM-L1$ algorithm, compute the components of the vectors of membership degrees: $\mathbf{u}_k^{(0)}=(u_{k1}^{(0)},\ldots,u_{kc}^{(0)})(k=1,\ldots,n)$ according to equation (7).
For $IFCM-L1$ algorithm, compute $J^{(t)}$ according to equation (1).
For $AIFCM-L1$ algorithm, compute $J^{(t)}$ according to equation(3)

5) *Stopping criterion*
If $|J^{(t)}-J^{(t-1)}|<\epsilon$ or $t>T$ then STOP; otherwise go to 2 (Representation step).

III. Experimental results

To evaluate the performance of these fuzzy c-means clustering algorithms, synthetic interval-valued data sets, a car

model and a freshwater fish species interval-valued data set are considered. Our aim is to achieve a comparison of the fuzzy clustering algorithms based on City-Block distances($IFCM-L1$ and $AIFCM-L1$) with the fuzzy clustering algorithms based on Euclidean distances($IFCM-L2$ and $AIFCM-L2$) [13] between vectors of intervals. Then, the usefulness of these fuzzy clustering algorithms will be illustrated with an application concerning a city temperatures interval-valued data set.

Intitially, each interval-valued variable on these data sets were normalized by means of a suitable dispersion measure.

Let $D_j = \{x_{1j}, \ldots, x_{nj}\}$ be the set of observed intervals $x_{ij} = [a_{ij}, b_{ij}]$ on variable $j(j = 1, \ldots, p)$. The dispersion of the jth variable is defined as [16] $s_j^2 = \sum_{i=1}^{n} d_j(x_{ij}, g_j)$, where $g_j = [\alpha_j, \beta_j]$ is the "central" interval computed from D_j and $d_j(x_{ij}, g_j) = (a_{ij} - \alpha_j)^2 + (b_{ij} - \beta_j)^2$ (if the comparison between intervals uses the Euclidean distance) or $d_j(x_{ij}, g_j) = |a_{ij} - \alpha_j| + |b_{ij} - \beta_j|$ (if the comparison between intervals uses the City-Block distance).

The central interval $g_j = (\alpha_j, \beta_j)^T$ has its bounds computed from $\sum_{i=1}^{n} d_j(x_{ij}, g_j) \rightarrow Min$.

$\tilde{z}_1, \ldots, \tilde{z}_n$

Each observed interval $x_{ij} = [a_{ij}, b_{ij}](i = 1, \ldots, n)$ is normalized as $\tilde{x}_{ij} = [\tilde{a}_{ij}, \tilde{b}_{ij}]$, where $\tilde{a}_{ij} = \frac{a_{ij}}{\sqrt{(s_j^2)}}$ and $\tilde{b}_{ij} = \frac{b_{ij}}{\sqrt{(s_j^2)}}$ (if the comparison between intervals uses the Euclidean distance) or $\tilde{a}_{ij} = \frac{a_{ij}}{s_j^2}$ and $\tilde{b}_{ij} = \frac{b_{ij}}{s_j^2}$ (if the comparison between intervals uses the Hausdorff distance). One can easily show that $\tilde{s}_j^2$ = 1, $\tilde{\alpha}_j = \frac{1}{n}\tilde{a}_{ij}$ and $\tilde{\beta}_j = \frac{1}{n}\tilde{b}_{ij}$. Now, all the normalized interval-valued variables have the same dispersion $\tilde{s}_j^2$ = 1.

Concerning these interval-valued data sets, each fuzzy clustering algorithm is run (until the convergence to a stationary value of the adequacy criterion) 100 times and the best result, according to the adequacy criterion, is selected. The parameter m was set at 2. Each fuzzy clustering algorithm gives a fuzzy partition. A hard partition $Q = (Q_1, \ldots, Q_K)$ is then obtained from this fuzzy partition $P = \{P_1, \ldots, P_K\}$, defining the cluster $Q_k(k = 1, \ldots, K)$ as: $Q_k = \{i \in \{1, \ldots, n\} : u_{ik} \geq u_{im}, \forall m \in \{1, \ldots, K\}\}$.

A. Synthetic interval-valued data sets

Each synthetic data set was created having classes with different sizes and shapes. The synthetic data sets has each 450 points divided into four classes of unequal sizes: two classes with ellipsis shape of size 150 each and two classes with spherical shapes of sizes 50 and 100. These data were drawn according to a bi-variate normal distribution with vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Data-set 1 gives well-separated classes. The data points of each class in this data set were drawn according to the following parameters:

a) Class 1: μ_1 = 28, μ_2 = 22, μ_1^2 = 100, μ_2^2 = 9, ρ = 0.7;

b) Class 2: μ_1 = 65, μ_2 = 30, μ_1^2 = 9, μ_2^2 = 144, ρ = 0.8;

c) Class 3: μ_1 = 45, μ_2 = 42, μ_1^2 = 9, μ_2^2 = 9, ρ = 0.7;

d) Class 4: μ_1 = 38, μ_2 = ..1, μ_1^2 = 25, μ_2^2 = 25, ρ = 0.8.

Data-set 2 gives overlapping classes. The data points of each class in this second data set were drawn according to the following parameters:

a) Class 1: μ_1 = 45, μ_2 = 22, μ_1^2 = 100, μ_2^2 = 9, ρ = 0.7;

b) Class 2: μ_1 = 65, μ_2 = 30, μ_1^2 = 9, μ_2^2 = 144, ρ = 0.8;

c) Class 3: μ_1 = 57, μ_2 = 42, μ_1^2 = 9, μ_2^2 = 9, ρ = 0.7;

d) Class 4: μ_1 = 42, μ_2 = -1, μ_1^2 = 25, μ_2^2 = 25, ρ = 0.8.

In order to build interval-valued data sets 1 and 2, each point (z_1, z_2) of these standard data sets is considered as the seed of a rectangle. Each rectangle is therefore a vector of two intervals defined by:

$([z_1 - \Im_1/2, z_1 + \Im_1/2], [z_2 - \Im_2/2, z_2 + \Im_2/2])$

The parameters $\Im_1$ and $\Im_2$ are the width and the height of the rectangle. They are drawn randomly within a given range of values. For example, the width and the height of all the rectangles can be drawn randomly within the interval [1, 8].

In order to compare the clustering results given by the adaptive and non-adaptive fuzzy c-means clustering algorithms for interval data $IFCM-L1$, $IFCM-L2$, $AIFCM-L1$, and $AIFCM-L2$, an external index, the corrected Rand index (CR), as well as the overall error rate of classification ($OERC$) will be considered.

For the synthetic data sets, the average of the CR and $OERC$ indexes are estimated in the framework of a Monte Carlo simulation with 75 replications for each data set. The average and standard deviation of these indexes among these 75 replications is calculated. In each replication a fuzzy clustering method is run (until the convergence to a stationary value of the adequacy criterion) 75 times and the best result, according to the adequacy criterion, is selected.

Table I shows the values of the average and the standard deviation (in parenthesis) of the CR and $OERC$ indexes for the different methods and for the data set 1 (well separated classes).

TABLE I. COMPARISON BETWEEN THE FUZZY C-MEANS CLUSTERING ALGORITHMS: AVERAGE AND STANDARD DEVIATION OF THE CR AND $OERC$ INDEXES

Interval-valued data set 1 (well-separated classes)				
Predefined intervals	Clustering algorithms			
	$IFCM-L1$		$AIFCM-L1$	
	CR	$OERC$	CR	$OERC$
[1,8]	0.9562 (0.0260)	0.0357 (0.0255)	0.9750 (0.0088)	0.0223 (0.0079)
[1,16]	0.9601 (0.0235)	0.0320 (0.0226)	0.9763 (0.0075)	0.0213 (0.0070)
[1, 24]	0.9600 (0.0222)	0.0323 (0.0215)	0.9759 (0.0094)	0.0214 (0.0086)
Predefined intervals	$IFCM-L2$		$AIFCM-L2$	
	CR	$OERC$	CR	$OERC$
[1,8]	0.9576 (0.0110)	0.0383 (0.0113)	0.9654 (0.0112)	0.0344 (0.0122)
[1,16]	0.9562 (0.0161)	0.0389 (0.0155)	0.9651 (0.0105)	0.0344 (0.0113)
[1,24]	0.9556 (0.0107)	0.0396 (0.0104)	0.9652 (0.0104)	0.0344 (0.0110)

It can be observed that the superiority of $AIFCM-L1$ in comparison with all the others fuzzy c-means clustering

algorithm is statiscally significant at the level of 5% (standard Student's t-test). Moreover, the superiority of $AIFCM-L2$ in comparison with $IFCM-L2$ is statiscally significant at the level of 5% only for CR index. Finally, the observed performance difference between $IFCM-L1$ and $IFCM-L2$ is not statiscally significant at the level of 5%.

Table II shows the values of the mean and the standard deviation of the CR index for the different methods and for the data set 2 (overlapping classes).

TABLE II. COMPARISON BETWEEN THE FUZZY C-MEANS CLUSTERING ALGORITHMS: AVERAGE AND STANDARD DEVIATION OF THE CR AND $OERC$ INDEXES

Interval-valued data set 2 (overlapping classes)				
	Clustering algorithms			
Predefined intervals	$IFCM-L1$		$AIFCM-L1$	
	CR	$OERC$	CR	$OERC$
[1,8]	0.8146 (0.0220)	0.1796 (0.0223)	0.8514 (0.0309)	0.1412 (0.0317)
[1,16]	0.8159 (0.0239)	0.1784 (0.0245)	0.8522 (0.0292)	0.1406 (0.0312)
[1, 24]	0.8146 (0.0222)	0.1797 (0.0213)	0.8442 (0.0334)	0.1480 (0.0355)
Predefined intervals	$IFCM-L2$		$AIFCM-L2$	
	CR	$OERC$	CR	$OERC$
[1,8]	0.7810 (0.0192)	0.2152 (0.0192)	0.8582 (0.0219)	0.1430 (0.0244)
[1,16]	0.7772 (0.0191)	0.2176 (0.0195)	0.8534 (0.0173)	0.1493 (0.0181)
[1,24]	0.7785 (0.0193)	0.2167 (0.0186)	0.8526 (0.0208)	0.1491 (0.0228)

It can be observed that the superiority of the adaptive methods ($AIFCM-L1$ and $AIFCM-L2$) in comparison with the non adaptive methods ($IFCM-L1$ and $IFCM-L2$) is statiscally significant at the level of 5%. Moreover, the observed performance difference between $AIFCM-L1$ and $AIFCM-L2$ is not statiscally significant at the level of 5%. Finally, the superiority of $IFCM-L1$ in comparison with $IFCM-L2$ is statiscally significant at the level of 5%.

B. Real interval-valued datasets

The car model data set concerns 33 car models described by 8 interval-valued variables. These car models are grouped in four a priori classes of unequal sizes: *Utilitarian* (size 10), *Berlina* (size 8),*Sporting* (size 7) and *Luxury* (size 8). The interval-valued variables are: *Price, Engine Capacity, Top Speed, Acceleration, Step, Length, Width* and *Height*.

The freshwater fish species concerns 12 species of freshwater fish, each specie being described by 13 symbolic interval variables. These species are grouped into four a *priori* classes of unequal sizes acoording to diet: two classes (*Carnivorous and Detritivorous*) of size 4 and two clusters of size 2 (*Omnivorous and Herbivorous*). The symbolic interval variables are:*Length, Weight, Muscle, Intestine, Stomach, Gills, Liver, Kidneys, Liver/Muscle, Kidneys/Muscle, Gills/Muscle, Intestine/Muscle* and *Stomach/Muscle*.

1) Performance of the fuzzy clustering algorithms: $IFCM-L1$ and $AIFCM-L1$ fuzzy clustering algorithms as well as $IFCM-L2$ and $AIFCM-L2$ fuzzy clustering algorithms [13] were performed on these data sets. From the 4-cluster fuzzy partitions furnished by these clustering algorithms, 4-cluster hard partitions were obtained. The 4-cluster hard partitions obtained with these clustering methods were compared with the 4-cluster partition known a *priori*. The comparison indexes used were the CR [17] and $OERC$ indexes.

Table III shows these indexes computed to the best run (the run that presented the minimum value for the adequacy criterion).

TABLE III. COMPARISON BETWEEN THE FUZZY CLUSTERING ALGORITHMS: THE INDEXES WERE COMPUTED FOR THE BEST RUN

Car models data set)		
	Comparison indexes	
Clustering alforithms	CR	$OERC$
$IFCM-L1$	0.3364	0.3333
$IFCM-L2$	0.2542	0.4545
$AIFCM-L1$	0.4998	0.2121
$AIFCM-L2$	0.5257	0.2121
Freshwater fish species data ser)		
	Comparison indexes	
Clustering alforithms	CR	$OERC$
$IFCM-L1$	0.2275	0.3333
$IFCM-L2$	0.0210	0.5000
$AIFCM-L1$	0.3473	0.3333
$AIFCM-L2$	0.2087	0.3333

For the car models data set, the fuzzy clustering algorithms with adaptive distances performed better than the fuzzy clustering algorithms with non-adaptive distances. Moreover, the non-adaptive fuzzy clustering algorithm performed better with the City-Block distances than with the Euclidean distances whereas the adaptive fuzzy clustering algorithm performed better with the Euclidean distances than with the City-Block distances. Finally, the worst performance was presented by the fuzzy clustering algorithm based on non-adaptive Euclidean distances.

For the freshwater fish species data set, the fuzzy clustering lgorithms with City-Block distances performed better than the fuzzy clustering algorithms with Euclidean distances. Moreover, the fuzzy clustering algorithms with non-adaptive Euclidean distances had a very poor performance.

2) Robustness of the fuzzy clustering algorithms: Theoretical studies indicate that city-block based models are more robust than those based on Euclidean distances. This point is adressed here. In order to evaluate the robustness of these fuzzy clustering algorithms, we introduces 4 outliers on the car models and freshwater fish species data sets in the following way: the boundaries of the interval-valued variables describing 4 individuals of these data sets (one individual by each a priori class) were multiplied by 10 (configuration 1), by 100 (configuration 2) and by 1000 (configuration 3).

$IFCM-L1$ and $AIFCM-L1$ as well as $IFCM-L2$ and $AIFCM-L2$ fuzzy clustering algorithms have been applied on these modified interval-valued data sets. From the 4-cluster fuzzy partitions furnished by these clustering algorithms, 4- cluster hard partitions were obtained.

The 4-cluster hard partitions obtained with these clustering methods were compared with the 4-cluster partition known a *priori*. Again, the comparison indexes used were the CR and the $OERC$ indexes. These indexes were also calculated for the best run. Table IV shows the results.

All the partitioning fuzzy clustering algorithms were affected by the introduction of outliers. However, those based

TABLE IV. COMPARISON BETWEEN THE FUZZY CLUSTERING ALGORITHMS: THE INDEXES WERE COMPUTED FOR THE BEST RUN

Car models interval-valued data set				
	Clustering algorithms			
	$IFCM-L1$		$AIFCM-L1$	
Configurations	CR	$OERC$	CR	$OERC$
1	0.2308	0.4848	0.2896	0.4545
2	0.2308	0.4848	0.2481	0.4848
3	0.2708	0.4848	0.2533	0.4848
Predefined	$IFCM-L2$		$AIFCM-L2$	
Configurations	CR	$OERC$	CR	$OERC$
1	0.2908	0.4848	0.3003	0.4848
2	-0.006	0.6363	-0.006	0.6363
3	-0.006	0.6363	-0.006	0.6363

Freshwater fish species interval-valued data set				
	Clustering algorithms			
	$IFCM-L1$		$AIFCM-L1$	
Configurations	CR	$OERC$	CR	$OERC$
1	0.0844	0.4166	0.3473	0.3333
2	0.1417	0.4166	0.3473	0.33338
3	-0.013	0.5000	0.3473	0.3333
Predefined	$IFCM-L2$		$AIFCM-L2$	
Configurations	CR	$OERC$	CR	$OERC$
1	0.1496	0.4166	0.2275	0.3333
2	-0.039	0.5000	0.2275	0.3333
3	0.0924	0.4848	0.2275	0.3333

on Euclidean distance were more affected than those based on City-Block distances.

IV. CONCLUSION

The main contributions of this paper were the introduction of partitioning fuzzy c-means clustering algorithms based on adaptive and non-adaptive City-Block distances for intervalvalued data. These fuzzy c-means clustering algorithms start from an initial fuzzy partition and alternate two steps (representation and allocation) or three steps (representation, weighting and allocation) respectively, for non-adaptive and adaptive City-Block distances, until the convergence of the algorithm, when the adequacy criterion reaches a stationary value. The performance of these fuzzy c-means clustering algorithms were evaluated in comparison with fuzzy c-means clustering algorithms based on Euclidean distances on synthetic real interval-valued data sets.

For the synthetic interval-valued data sets showing well separated a priori classes, the fuzzy c-means clustering algorithm based on adaptive City-Block distances presented the best performance. For the interval-valued data sets showing overlapping a priori classes, the adaptive methods ($AIFCM-L1$ and $AIFCM-L2$) were superior the non adaptive methods ($IFCM-L1$ and $IFCM-L2$).

Concerning the freshwater fish species and car models interval-valued data sets, the fuzzy c-means clustering algorithms with adaptive distances outperformed the fuzzy c-means clustering algorithms with non-adaptive distances. Moreover, the non-adaptive fuzzy c-means clustering algorithm performed better with the City-Block distances than with the Euclidean distances, whereas the adaptive fuzzy c-means clustering algorithm performed better with the Euclidean distances on the car models data set and with the City-Block distances on the freshwater fish species data set. Finally, all the partitioning fuzzy c-means clustering algorithms were affected by the introduction of outliers. However, those based on Euclidean distance were more affected than those based on City-Block distances.

ACKNOWLEDGMENT

The authors would like to thank CNPq and FACEPE (Brazilian agencies) for their financial support and the anonymous referees for their helpful comments and suggestions for improving the paper.

REFERENCES

[1] H.-H. Bock and E. Diday, *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data.* Berlin: Springer, 2000.

[2] E. Diday and M. Noirhomme-Fraiture, *Symbolic Data Analysis and the Sodas Software.* Chichester: Wiley, 2008.

[3] A. D. Gordon, *Classification.* Boca Raton, Florida: Chapman and Hall/CRC, 1999.

[4] B. Everitt, *Cluster Analysis.* Halsted, New York: Chapman and Hall/CRC, 2001.

[5] A. K. Jain, "Data clustering: 50 years beyond k-means,"*Pattern Recognition Letters*, vol. 31, pp. 651666, 2010.

[6] J. C. Dunn, "A fuzzy relative to the isodata process and its use in detecting compact, well-separated clusters," *J. Cybernet.*, vol. 3, pp. 3257, 1974.

[7] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms.* New York: Plenum, 1981.

[8] E. Diday and G. Govaert, "Classification automatique avec distances adaptatives," R.A.I.R.O. *Informatique Computer Science*, vol. 1, no. 4, pp. 329349, 1977.

[9] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *Proc. IEEE Conf. Decision Contr.*, 1979, pp. 761766.

[10] F. A. T. De Carvalho, C. P. Tenório, and N. L. Cavalcanti Junior, "Partitional fuzzy clustering methods based on adaptive quadratic distances," *Fuzzy Sets and Systems*, vol. 157, pp. 28332857, 2006.

[11] Y. El-Sonbaty and M. A. Ismail, "Fuzzy clustering for symbolic data,"*IEEE Transactions on Fuzzy Systems,* vol. 6, pp. 195204, 1998.

[12] M.-S. Yang, P.-Y. Hwang, and D.-H. Chen, "Fuzzy clustering algorithms for mixed feature variables," *Fuzzy Sets and Systems,* vol. 141, pp. 301317, 2004.

[13] F. A. T. De Carvalho, "Fuzzy c-means clustering methods for symbolic interval data,"*Pattern Recognition Letters,* vol. 28, no. 4, pp. 423437,2007.

[14] F. A. T. De Carvalho and C. P. Tenório, "Fuzzy k-means clustering algorithms for interval-valued data based on adaptive quadratic distances."*Fuzzy Sets and Systems,* vol. 161, pp. 29782999, 2010.

[15] K. Jajuga, "L1-norm based fuzzy clustering," *Fuzzy Sets and Systems,* vol. 39, pp. 4350, 1991.

[16] M. Chavent and J. Saracco, "On central tendency and dispersion measures for intervals and hypercubes," *Communications in Statistics Theory and Methods,* vol. 37, pp. 14711482, 2008.

[17] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification,* vol. 2, pp. 193218, 1985.