Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition Letters xxx (2006) xxx–xxx

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Fuzzy *c*-means clustering methods for symbolic interval data

Francisco de A.T. de Carvalho *

Centro de Informática, Universidade Federal de Pernambuco, Caixa Postal 7851, CEP 50732-970 Recife (PE), Brazil

Received 17 August 2005; received in revised form 11 August 2006

Communicated by W. Pedrycz

Abstract

This paper presents adaptive and non-adaptive fuzzy *c*-means clustering methods for partitioning symbolic interval data. The proposed methods furnish a fuzzy partition and prototype for each cluster by optimizing an adequacy criterion based on suitable squared Euclidean distances between vectors of intervals. Moreover, various cluster interpretation tools are introduced. Experiments with real and synthetic data sets show the usefulness of these fuzzy *c*-means clustering methods and the merit of the cluster interpretation tools.
© 2006 Published by Elsevier B.V.

Keywords: Symbolic data analysis; Fuzzy *c*-means clustering methods; Symbolic interval data; Squared euclidean distances; Adaptive distances; Fuzzy partition interpretation indices; Fuzzy cluster interpretation indices

1. Introduction

Clustering methods seeks to organize a set of items into clusters such that items within a given cluster have a high degree of similarity, whereas items belonging to different clusters have a high degree of dissimilarity. These methods have been widely applied in various areas such as taxonomy, image processing, information retrieval, data mining, etc. Clustering techniques may be divided into hierarchical and partitioning methods (Jain et al., 1999; Gordon, 1999): hierarchical methods yield 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 in a fixed number of clusters, usually by optimizing an objective function.

In clustering analysis, the patterns to be grouped are usually represented as a vector of quantitative or qualitative measurements where each column represents a variable. Each pattern takes a single value for each variable. However, this model is too restrictive to represent complex data. In order to take into account variability and/or uncertainty inherent to the data, variables must assume sets of categories or intervals, possibly even with frequencies or weights. These kinds of data have been mainly studied in *Symbolic Data Analysis* (SDA), a new domain related to multivariate analysis, pattern recognition and artificial intelligence. The aim of Symbolic Data Analysis is to provide suitable methods (clustering, factorial techniques, decision trees, etc.) for managing aggregated data described by multi-valued variables, where the cells of the data table contain sets of categories, intervals, or weight (probability) distributions (Bock and Diday, 2000; Billard and Diday, 2003).

SDA provides a number of clustering methods for symbolic data. These methods differ in the type of the considered symbolic data, in their cluster structures and/or in the considered clustering criteria. With hierarchical methods, an agglomerative approach has been introduced that forms composite symbolic objects using a join operator whenever mutual pairs of symbolic objects are selected for agglomeration based on minimum dissimilarity (Gowda and Diday, 1991) or maximum similarity (Gowda and Diday, 1992). Ichino and Yaguchi (1994) defined generalized Minkowski metrics for mixed feature variables and presents dendrograms obtained from the application of

* Tel.: +55 81 21268430; fax: +55 81 21268438.
E-mail address: fatc@cin.utpe.br

0167-8655/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.patrec.2006.08.014

standard linkage methods for data sets containing numeric and symbolic feature values. Gowda and Ravi (1995a) and Gowda and Ravi (1995b), respectively, presented divisive and agglomerative algorithms for symbolic data based on the combined usage of similarity and dissimilarity measures. These proximity measures are defined on the basis of the position, span and content of symbolic data. Chavent (1998) proposed a divisive clustering method that simultaneously furnishes a hierarchy of the symbolic data set and a monothetic characterization of each cluster in the hierarchy. Gowda and Ravi (1999a) presented a hierarchical clustering algorithm for symbolic data based on the gravitational approach, which is inspired on the movement of particles in space due to their mutual gravitational attraction. Guru et al. (2004) and Guru and Kiranagi (2005) introduced agglomerative clustering algorithms based, respectively, on similarity and dissimilarity functions that are multi-valued and non-symetric.

A number of 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 manage data characterized by numerical and categorical variables, and complemented this method with a characterization algorithm to provide a conceptual interpretation of the resulting clusters. 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 has 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. Souza and De Carvalho (2004) proposed partitioning clustering methods for interval data based on city-block distances, also considering adaptive distances. More recently, De Carvalho et al. (2006) proposed an algorithm using an adequacy criterion based on adaptive Hausdorff distances.

Conventional hard clustering methods restrict each point of the data set to exactly one cluster. Fuzzy clustering generates a fuzzy partition based on the idea of partial membership expressed by the degree of membership of each pattern in a given cluster. Concerning quantitative data, Dunn (1974) presented one of the first fuzzy clustering methods based on an adequacy criterion defined by the Euclidean distance. Bezdek (1981) further generalized this method. Diday and Govaert (1977) introduced one of the first approaches to use adaptive distances in partitioning quantitative data. Gustafson and Kessel (1979) introduced the first adaptive fuzzy clustering algorithm, based on a quadratic distance defined by a fuzzy covariance matrix. El-Sonbaty and Ismail (1998) presented a fuzzy c-means algorithm to cluster data on the basis of different types of symbolic variables. Yang et al. (2004) presented fuzzy clustering algorithms for mixed features of symbolic and fuzzy data. In these fuzzy clustering algorithms, the membership degree is associated to the values of the features in the clusters for the cluster centers instead of being associated to the patterns in each cluster, as is the usual case.

As pointed out above, items to be clustered are usually represented as a vector of quantitative measurements. However, due to recent advances in database technologies, it is now common to record interval data. Therefore, tools for symbolic interval data analysis are very much required. This paper introduces adaptive and non-adaptive fuzzy c-means clustering algorithms for symbolic interval data, as well as various tools for fuzzy partition and cluster interpretation suitable for these fuzzy clustering algorithms.

Section 2 presents the (adaptive and non-adaptive) fuzzy c-means clustering algorithms for partitioning symbolic interval data. Celeux et al. (1989) introduced a family of indices for interpreting a hard partition of classical quantitative data based on the notion of the sum of squares. In this paper, we adapt these indices to fuzzy partitions of symbolic interval data. In Section 3, we propose various tools for cluster interpretation according to the different fuzzy clustering models: 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. To show the usefulness of these fuzzy clustering algorithms and the merit of these cluster interpretation tools, experiments with simulated data in a framework of a Monte Carlo schema as well as applications with real symbolic interval data sets are considered (Section 4). Section 5 gives the concluding remarks.

2. Fuzzy *c*-means clustering methods for symbolic interval data

This section introduces two fuzzy c-means clustering methods for symbolic interval data. The first method is a suitable extension of the standard fuzzy c-means clustering algorithm that furnishes a fuzzy partition and a prototype for each cluster by optimizing an adequacy criterion based on a suitable squared Euclidean distance between vectors of intervals. The second method introduces an adaptive version of the the first method, where the adequacy criterion is based on a suitable adaptive squared Euclidean distance.

2.1. Fuzzy c-means clustering method for symbolic interval data

Let $\Omega = \{1, \ldots, n\}$ be a set of n patterns (each pattern is indexed by k) described by p symbolic interval variables

$\{y_1, \ldots, y_p\}$ (each variable is indexed by j). A *symbolic interval variable X* (Bock and Diday, 2000) is a correspondence defined from Ω in $\Re$ such that for each $k \in \Omega, X(k) = [a, b] \in \Im$, where $\Im = \{[a, b] : a, b \in \Re, \ a \leqslant b\}$ is the set of closed intervals defined from $\Re$. Each pattern k is represented as a vector of intervals $\boldsymbol{x}_k = (x_k^1, \cdots, x_k^p)$, where $x_k^j = [a_k^j, b_k^j] \in \Im$. In this paper, an interval data table $\{x_k^j\}_{n \times p}$ is made up of n rows representing the n patterns to be clustered, and p columns representing p symbolic interval variables. Each cell of this table contains an interval $x_k^j = [a_k^j, b_k^j] \in \Im$. Let each prototype $\boldsymbol{g}_i$ of cluster P_i be also represented as a vector of intervals $\boldsymbol{g}_i = (g_i^1, \cdots, g_i^p)$, where $g_i^j = [\alpha_i^j, \beta_i^j] \in \Im$.

As in the standard fuzzy c-means algorithm (Bezdek, 1981), the fuzzy c-means clustering method for symbolic interval data (here labeled IFCM) aims to furnish a fuzzy partition of a set of patterns in c clusters $\{P_1, \ldots, P_c\}$ and a corresponding set of prototypes $\{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_c\}$ such that a criterion W^1 measuring the fitting between the clusters and their representatives (prototypes) is locally minimized. This criterion is based on a non-adaptive squared Euclidean distance between vectors of intervals and is defined as:

$$\begin{aligned} W^1 &= \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \phi(\boldsymbol{x}_k, \boldsymbol{g}_i) \\ &= \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} [(a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2] \end{aligned} \tag{1}$$

where ϕ is the square of a suitable Euclidean distance measuring the dissimilarity between a pair of vectors of intervals, $\boldsymbol{x}_k = (x_k^1, \ldots, x_k^p)$ is a vector of intervals describing the kth pattern, $\boldsymbol{g}_i = (g_i^1, \ldots, g_i^p)$ is a vector of intervals describing the prototype of class P_i, u_{ik} is the membership degree of pattern k in cluster P_i and $m \in]1, +\infty[$ is a parameter that controls the fuzziness of membership for each pattern k.

As in the standard fuzzy c-means algorithm (Bezdek, 1981), this algorithm sets an initial membership degree for each pattern k in each cluster P_i and alternates a *representation step* and an *allocation step* until convergence when the criterion W^1 reaches a stationary value representing a local minimum.

2.1.1. Representation step: definition of the best prototypes

In the representation step, the membership degree u_{ik} of each pattern k in cluster P_i is fixed.

Proposition 2.1. *The prototype* $\boldsymbol{g}_i = (g_i^1, \ldots, g_i^p)$ *of class* P_i $(i = 1, \ldots, c)$, *which minimizes the clustering criterion* W^1, *has the bounds of the interval* $g_i^j = [\alpha_i^j, \beta_i^j]$ $(j = 1, \ldots, p)$ *updated according to the following expression*:

$$\alpha_i^j = \frac{\sum_{k=1}^{n} (u_{ik})^m a_k^j}{\sum_{k=1}^{n} (u_{ik})^m} \quad \text{and} \quad \beta_i^j = \frac{\sum_{k=1}^{n} (u_{ik})^m b_k^j}{\sum_{k=1}^{n} (u_{ik})^m}, \quad \text{for } j = 1, \ldots, p \tag{2}$$

Proof. The proof can be obtained in a similar way as described in Bezdek (1981) for the case of standard quantitative data. □

2.1.2. Allocation step: definition of the best fuzzy partition

In the allocation step, each prototype $\boldsymbol{g}_i$ of class P_i $(i = 1, \ldots, c)$ is fixed.

Proposition 2.2. *The membership degree* u_{ik} $(k = 1, \ldots, n)$ *of each pattern k in each cluster* P_i, *minimizing the clustering criterion* W^1 *under* $u_{ik} \geqslant 0$ *and* $\sum_{i=1}^{c} u_{ik} = 1$, *is updated according to the following expression*:

$$u_{ik} = \left[\sum_{h=1}^{c} \left(\frac{\sum_{j=1}^{p} [(a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2]}{\sum_{j=1}^{p} [(a_k^j - \alpha_h^j)^2 + (b_k^j - \beta_h^j)^2]} \right)^{\frac{1}{m-1}} \right]^{-1} \quad \text{for } i = 1, \ldots, c \tag{3}$$

Proof. The proof can be obtained in a similar way as described in Bezdek (1981) for the case of standard quantitative data. □

2.1.3. Algorithm

The IFCM clustering algorithm for symbolic interval data is executed in the following steps:

```
(1) Initialization
    Fix $c$, $2 \leqslant c < n$; fix $m$, $1 < m < \infty$; fix $T$ (an iteration limit); and fix $\varepsilon > 0$; Initialize $u_{ik}$ ($k = 1, \ldots, n$ and $i = 1, \ldots, c$) of pattern $k$ belonging to cluster $P_i$ such that $u_{ik} \geqslant 0$ and $\sum_{i=1}^{c} u_{ik} = 1$
(2) $t = 1$
(3) Representation step:
    {the membership degree $u_{ik}$ of pattern $k$ belonging to cluster $P_i$ is fixed}
    Compute the prototypes $\boldsymbol{g}_i$ of class $P_i$ ($i = 1, \ldots, c$) using Eq. (2)
(4) Allocation step:
    {the prototypes $\boldsymbol{g}_i$ of class $P_i$ ($i = 1, \ldots, c$) are fixed}
    Update the fuzzy membership degree $u_{ik}$ of pattern $k$ belonging to cluster $P_i$ ($i = 1, \ldots, c$) using Eq. (3)
(5) Stopping criterion
    If $|W_{t+1}^1 - W_t^1| \leqslant \varepsilon$ or $t > T$
    stop
    else $t = t + 1$ and go to step 3
```

2.2. An adaptive fuzzy c-means clustering method for symbolic interval data

In this section, we present a fuzzy c-means clustering method for symbolic interval data based on an adaptive squared Euclidean distance between vectors of intervals (here labeled IFCMADC). The main idea is that there is a different distance associated to each cluster for comparing clusters and their representatives that changes at each iteration, i.e., the distance is not definitively determined and is different from one class to another. The advantage of these adaptive distances is that the clustering algorithm is able to find clusters of different shapes and sizes (Diday and Govaert, 1977; Gustafson and Kessel, 1979).

This adaptive method looks for a fuzzy partition of a set of patterns in c clusters $\{P_1,\ldots,P_c\}$, the corresponding c prototypes $\{\boldsymbol{g}_1,\ldots,\boldsymbol{g}_c\}$ and the square of an adaptive squared Euclidean distance between vectors of intervals that is different for each class, such that a criterion W^2 measuring the fitting between the clusters and their representatives (prototypes) is locally minimized. This criterion W^2 is based on an adaptive squared Euclidean distance for each cluster and is defined as

$$
\begin{aligned}
W^2 &= \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \psi_i(\boldsymbol{x}_k,\boldsymbol{g}_i) \\
&= \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \sum_{j=1}^{p}\lambda_i^j[(a_k^j-\alpha_i^j)^2+(b_k^j-\beta_i^j)^2] \qquad (4)
\end{aligned}
$$

where $\boldsymbol{x}_k$, $\boldsymbol{g}_i$, u_{ik} and m are defined as before and ψ is now the square of an adaptive Euclidean distance defined for each class and parameterized by the vectors of weights $\boldsymbol{\lambda}_i = (\lambda_i^1,\ldots,\lambda_i^p)(i=1,\ldots,c)$, which change at each iteration.

The algorithm starts from an initial membership degree for each pattern k in each cluster P_i and alternates a representation step and an allocation step until the convergence, when the criterion W^2 reaches a stationary value representing a local minimum. The representation step now has two stages.

2.2.1. Representation step: definition of the best prototypes

In the first stage, the membership degree u_{ik} of each pattern k in cluster P_i and the vectors of weights $\boldsymbol{\lambda}_i = (\lambda_i^1,\ldots,\lambda_i^p)$ $(i=1,\ldots,c)$ are fixed.

Proposition 2.3. *The prototype* $\boldsymbol{g}_i = (g_i^1,\ldots,g_i^p)$ *of class* P_i $(i = 1,\ldots,c)$, *which minimizes the clustering criterion* W^2, *has the bounds of the interval* $g_i^j = [\alpha_i^j,\beta_i^j]$ $(j=1,\ldots,p)$ *updated according to the following expression*:

$$
\alpha_i^j = \frac{\sum_{k=1}^{n}(u_{ik})^m a_k^j}{\sum_{k=1}^{n}(u_{ik})^m} \quad \text{and} \quad \beta_i^j = \frac{\sum_{k=1}^{n}(u_{ik})^m b_k^j}{\sum_{k=1}^{n}(u_{ik})^m} \quad \text{for } j=1,\ldots,p \qquad (5)
$$

Proof. The proof can be obtained in a similar way as described in Bezdek (1981) for the case of standard quantitative data. □

2.2.2. Representation step: definition of the best distances

In the second stage, the membership degree u_{ik} of each pattern k in cluster P_i and the prototypes $\boldsymbol{g}_i$ of class P_i $(i=1,\ldots,c)$ are fixed.

Proposition 2.4. *The vectors of weights* $\boldsymbol{\lambda}_i = (\lambda_i^1,\ldots,\lambda_i^p)$ $(i=1,\ldots,c)$, *which minimize the clustering criterion* W^2 *under* $\lambda_i^j > 0$ *and* $\prod_{j=1}^{p}\lambda_i^j = 1$, *are updated according to the following expression*:

$$
\lambda_i^j = \frac{\left\{\prod_{h=1}^{p}\left[\sum_{k=1}^{n}(u_{ik})^m((a_k^h-\alpha_i^h)^2+(b_k^h-\beta_i^h)^2)\right]\right\}^{\frac{1}{p}}}{\sum_{k=1}^{n}(u_{ik})^m[(a_k^j-\alpha_i^j)^2+(b_k^j-\beta_i^j)^2]}, \quad j=1,\ldots,p \qquad (6)
$$

Proof. The proof is given in Appendix A. □

2.2.3. Allocation step: definition of the best fuzzy partition

In the allocation step, the prototypes $\boldsymbol{g}_i$ of class P_i $(i=1,\ldots,c)$ and the vectors of weights $\boldsymbol{\lambda}_i = (\lambda_i^1,\ldots,\lambda_i^p)$ $(i=1,\ldots,c)$ are fixed.

Proposition 2.5. *The membership degree* u_{ik} $(k = 1,\ldots,n)$ *of each pattern* k *in each cluster* P_i, *minimizing the clustering criterion* W^2 *under* $u_{ik} \geqslant 0$ *and* $\sum_{i=1}^{c}u_{ik} = 1$, *is updated according to the following expression*:

$$
u_{ik} = \left[\sum_{h=1}^{c}\left(\frac{\sum_{j=1}^{p}\lambda_i^j[(a_k^j-\alpha_i^j)^2+(b_k^j-\beta_i^j)^2]}{\sum_{j=1}^{p}\lambda_h^j[(a_k^j-\alpha_h^j)^2+(b_k^j-\beta_h^j)^2]}\right)^{\frac{1}{m-1}}\right]^{-1} \quad (i=1,\ldots,c) \qquad (7)
$$

Proof. The proof can be obtained in a similar way as described in Bezdek (1981) for the case of standard quantitative data. □

2.2.4. Algorithm

The IFCMADC clustering algorithm is executed in the following steps:

(1) **Initialization**
Fix c, $2 \leqslant c < n$; fix m, $1 < m < \infty$; fix T (an iteration limit); and fix $\varepsilon > 0$; Initialize u_{ik} $(k=1,\ldots,n$ and $i=1,\ldots,c)$ of pattern k belonging to cluster P_i such that $u_{ik} \geqslant 0$ and $\sum_{i=1}^{c}u_{ik} = 1$

(2) $t = 1$

(3) **Representation step**:
 (a) *Stage 1*:
 {the membership degree u_{ik} of pattern k belonging to cluster P_i is fixed}
 Compute the prototypes $\boldsymbol{g}_i$ of class P_i $(i=1,\ldots,c)$ using Eq. (5)
 (b) *Stage 2*:
 {the membership degree u_{ik} of pattern k belonging to cluster P_i and the prototypes $\boldsymbol{g}_i$ of class P_i $(i=1,\ldots,c)$ are fixed}
 Compute the vector of weights $\boldsymbol{\lambda}_i$ for $i=1,\ldots,c$ using Eq. (6)

(4) **Allocation step**:
{the prototypes $\boldsymbol{g}_i$ of class P_i $(i=1,\ldots,c)$ and the vector of weights $\boldsymbol{\lambda}_i$ for $i=1,\ldots,c$ are fixed}
Update the fuzzy membership degree u_{ik} of pattern k belonging to cluster P_i $(i=1,\ldots,c)$ using Eq. (7)

Please cite this article in press as: de Carvalho, F.d.A.T., Fuzzy c-means clustering methods for symbolic interval data, Pattern Recog. Lett. (2006), doi:10.1016/j.patrec.2006.08.014

```
(5) Stopping criterion
    If $|W^2_{t+1} - W^2_t| \leqslant \varepsilon$ or $t > T$
    stop
    else $t = t + 1$ and go to step 3
```

3. Partition and cluster interpretation

Partition and cluster interpretation is an important step in clustering analysis. The user wants to evaluate the overall data heterogeneity, intra-cluster and between-cluster heterogeneity, the contribution of each variable to cluster formation, etc. For the case of usual quantitative data partitioned by the hard c-means clustering algorithm, Celeux et al. (1989) introduced a family of indices for cluster interpretation that are based on the sum of squares (SSQ). In this section, we adapt these indices to the case of symbolic interval data partitioned by the fuzzy c-means clustering algorithms presented in this paper.

We consider the fuzzy partition $\{P_1, \ldots, P_c\}$ of $\Omega = \{1, \ldots, n\}$ in c clusters that was obtained from one of the methods presented in Sections 2.1 and 2.2 and denoted by

$$\boldsymbol{g}_i = \{g_i^1, \ldots, g_i^p\}, g_i^j = [\alpha_i^j, \beta_i^j](j = 1, \ldots, p),$$

$$\text{with } \alpha_i^j = \frac{\sum_{k=1}^n (u_{ik})^m a_k^j}{\sum_{k=1}^n (u_{ik})^m} \text{ and } \beta_i^j = \frac{\sum_{k=1}^n (u_{ik})^m b_k^j}{\sum_{k=1}^n (u_{ik})^m}$$

$(j = 1, \ldots, p)$, the prototype of cluster P_i

Moreover, the vector $\boldsymbol{z} = (z^1, \ldots, z^p\}, z^j = [\alpha^j, \beta^j]$ $(j = 1, \ldots, p)$, with

$$\alpha^j = \frac{\sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m a_k^j}{\sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m} = \frac{\sum_{i=1}^c (\sum_{k=1}^n (u_{ik})^m \alpha_i^j)}{\sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m} = \frac{\sum_{i=1}^c \mu_i \alpha_i^j}{\sum_{i=1}^c \mu_i},$$

$$\beta^j = \frac{\sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m b_k^j}{\sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m} = \frac{\sum_{i=1}^c (\sum_{k=1}^n (u_{ik})^m \beta_i^j)}{\sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m} = \frac{\sum_{i=1}^c \mu_i \beta_i^j}{\sum_{i=1}^c \mu_i}$$

$$\text{and} \quad \mu_i = \sum_{k=1}^n (u_{ik})^m$$

$(j = 1, \ldots, p)$ is the overall representative vector for all n interval data points.

3.1. Measures based on the sum of squares

In this section, we define the overall SSQ and SSQ within and between clusters for symbolic interval data in the framework of the presented partitioning fuzzy c-means clustering algorithms, and we show that this overall sum of squares decomposes into the sum of squares within clusters plus the sum of squares between clusters. This decomposition is the basis for defining the interpretation tools in Section 3.2.

3.1.1. Overall fuzzy sum of squares

According to the distance function used, the overall heterogeneity of all n interval data patterns is measured by the overall fuzzy sum of squares

$$T^1 = \sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m \phi(\boldsymbol{x}_k, \boldsymbol{z}) = \sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m \sum_{j=1}^p [(a_k^j - \alpha^j)^2 + (b_k^j - \beta^j)^2]$$
$$T^2 = \sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m \psi_i(\boldsymbol{x}_k, \boldsymbol{z}) = \sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m \sum_{j=1}^p \lambda_i^j \left[(a_k^j - \alpha^j)^2 + (b_k^j - \beta^j)^2\right] \tag{8}$$

T^l $(l = 1, 2)$ decomposes, on the one hand, into the fuzzy sum of the cluster-specific SSQs in the clusters P_i given by $T^l = \sum_{i=1}^c T_i^l$ $(l = 1, 2)$ with

$$T_i^1 = \sum_{k=1}^n (u_{ik})^m \phi(\boldsymbol{x}_k, \boldsymbol{z}) = \sum_{k=1}^n (u_{ik})^m \sum_{j=1}^p [(a_k^j - \alpha^j)^2 + (b_k^j - \beta^j)^2]$$
$$T_i^2 = \sum_{k=1}^n (u_{ik})^m \psi_i(\boldsymbol{x}_k, \boldsymbol{z}) = \sum_{k=1}^n (u_{ik})^m \sum_{j=1}^p \lambda_i^j [(a_k^j - \alpha^j)^2 + (b_k^j - \beta^j)^2] \tag{9}$$

and, on the other hand, into the fuzzy sum of the variable-specific overall SSQs for the variables $j = 1, \ldots p$ given by $T^l = \sum_{j=1}^p T_j^l$ $(l = 1, 2)$ with:

$$T_j^1 = \sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m [(a_k^j - \alpha^j)^2 + (b_k^j - \beta^j)^2]$$
$$T_j^2 = \sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m \lambda_i^j [(a_k^j - \alpha^j)^2 + (b_k^j - \beta^j)^2] \tag{10}$$

In both cases, $T^l = \sum_{i=1}^c T_i^l = \sum_{i=1}^c \left(\sum_{j=1}^p T_{ij}^l\right)$ $(l = 1, 2)$ and

$$T_{ij}^1 = \sum_{k=1}^n (u_{ik})^m [(a_k^j - \alpha^j)^2 + (b_k^j - \beta^j)^2]$$
$$T_{ij}^2 = \sum_{k=1}^n (u_{ik})^m \lambda_i^j [(a_k^j - \alpha^j)^2 + (b_k^j - \beta^j)^2] \tag{11}$$

denote the partial SSQ in the class P_i relating to the jth variable $(j = 1, \ldots, p;\ i = 1, \ldots, c)$.

3.1.2. Within-cluster fuzzy sum of squares

Here we consider the heterogeneity within the clusters P_i and measure it by the within-cluster fuzzy SSQ according to the distance function used:

$$W_i^1 = \sum_{k=1}^n (u_{ik})^m \phi(\boldsymbol{x}_k, \boldsymbol{g}_i) = \sum_{k=1}^n (u_{ik})^m \sum_{j=1}^p [(a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2]$$
$$W_i^2 = \sum_{k=1}^n (u_{ik})^m \psi_i(\boldsymbol{x}_k, \boldsymbol{g}_i) = \sum_{k=1}^n (u_{ik})^m \sum_{j=1}^p \lambda_i^j [(a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2] \tag{12}$$

Summing up all clusters, we obtain the overall within-cluster fuzzy SSQ $W^l = \sum_{i=1}^c W_i^l$ $(l = 1, 2)$, i.e.,

$$W^1 = \sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m \phi(\boldsymbol{x}_k, \boldsymbol{g}_i) = \sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m \sum_{j=1}^p [(a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2]$$
$$W^2 = \sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m \psi_i(\boldsymbol{x}_k, \boldsymbol{g}_i) = \sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m \sum_{j=1}^p \lambda_i^j [(a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2] \tag{13}$$

On the other hand, W^l $(l = 1, 2)$ decomposes into the sum of the variable-specific overall within-cluster fuzzy SSQs

Please cite this article in press as: de Carvalho, F.d.A.T., Fuzzy c-means clustering methods for symbolic interval data, Pattern Recog. Lett. (2006), doi:10.1016/j.patrec.2006.08.014

for the variables $j = 1,\dots,p$, given by $W^l = \sum_{j=1}^{p} W_j^l$ $(l = 1,2)$, with

$$W_j^1 = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m[(a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2]$$
$$W_j^2 = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m\lambda_i^j[(a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2] \tag{14}$$

In both cases, $W^l = \sum_{i=1}^{c} W_i^l = \sum_{i=1}^{c}\left(\sum_{j=1}^{p} W_{ij}^l\right)$ $(l = 1,2)$ and

$$W_{ij}^1 = \sum_{k=1}^{n}(u_{ik})^m[(a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2]$$
$$W_{ij}^2 = \sum_{k=1}^{n}(u_{ik})^m\lambda_i^j[(a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2]$$

denote the within-cluster fuzzy SSQ W_{ij}^l $(l = 1,2)$ of the variable j in cluster P_i $(j = 1,\dots,p;\ i = 1,\dots,c)$.

3.1.3. Between-cluster fuzzy sum of squares

The between-cluster fuzzy SSQ given by

$$B^1 = \sum_{i=1}^{c}\mu_i\phi(\boldsymbol{g}_i,\boldsymbol{z}) = \sum_{i=1}^{c}\mu_i\sum_{j=1}^{p}[(\alpha_i^j - \alpha^j)^2 + (\beta_i^j - \beta^j)^2]$$
$$B^2 = \sum_{i=1}^{c}\mu_i\psi_i(\boldsymbol{g}_i,\boldsymbol{z}) = \sum_{i=1}^{c}\mu_i\sum_{j=1}^{p}\lambda_i^j[(\alpha_i^j - \alpha^j)^2 + (\beta_i^j - \beta^j)^2] \tag{15}$$

measures the dispersion of the cluster representatives and, consequently, the distinctness of all clusters. It is decomposed either into the sum of all c cluster-specific fuzzy SSQs $B^l = \sum_{i=1}^{c} B_i^l$ $(l = 1,2)$ with

$$B_i^1 = \sum_{j=1}^{p} B_{ij}^1 = \mu_i\phi(\boldsymbol{g}_i,\boldsymbol{z}) = \mu_i\sum_{j=1}^{p}[(\alpha_i^j - \alpha^j)^2 + (\beta_i^j - \beta^j)^2]$$
$$B_i^2 = \sum_{j=1}^{p} B_{ij}^2 = \mu_i\psi_i(\boldsymbol{g}_i,\boldsymbol{z}) = \mu_i\sum_{j=1}^{p}\lambda_i^j[(\alpha_i^j - \alpha^j)^2 + (\beta_i^j - \beta^j)^2] \tag{16}$$

which measures the heterogeneity in the clusters P_i, or into the sum of the p variable-specific between-cluster fuzzy SSQs $B^l = \sum_{j=1}^{p} B_j^l$ $(l = 1,2)$ as given by

$$B_j^1 = \sum_{i=1}^{c} B_{ij}^1 = \sum_{i=1}^{c}\mu_i[(\alpha_i^j - \alpha^j)^2 + (\beta_i^j - \beta^j)^2]$$
$$B_j^2 = \sum_{i=1}^{c} B_{ij}^3 = \sum_{i=1}^{c}\mu_i\lambda_i^j[(\alpha_i^j - \alpha^j)^2 + (\beta_i^j - \beta^j)^2] \tag{17}$$

In all formulas

$$B_{ij}^1 = \mu_i[(\alpha_i^j - \alpha^j)^2 + (\beta_i^j - \beta^j)^2]$$
$$B_{ij}^2 = \mu_i\lambda_i^j[(\alpha_i^j - \alpha^j)^2 + (\beta_i^j - \beta^j)^2] \tag{18}$$

measures the dissimilarity (for the variable j) between the cluster prototype $\boldsymbol{g}_i$ of P_i and the overall prototype $\boldsymbol{z}$ of all data, with a factor μ_i for considering all cluster members.

The following result establishes that the overall fuzzy sum of squares decomposes into the fuzzy sum of squares within clusters plus the fuzzy sum of squares between clusters.

Proposition 3.1. *For* $l = 1,2$; $i = 1,\dots,c$; $j = 1,\dots,p$, *the following relations hold*:

$$T^l = W^l + B^l,\quad T_i^l = B_i^l + W_i^l,\quad T_j^l = B_j^l + W_j^l,$$
$$T_{ij}^l = B_{ij}^l + W_{ij}^l \tag{19}$$

Proof. The proof is given in Appendix A. □

3.2. Interpretation indices

The indices introduced in this section are a suitable adaptation of the indices presented in (Celeux et al., 1989) for the case of the hard c-means clustering algorithm for quantitative data. All these indices range between 0 and 1.

3.2.1. Fuzzy partition interpretation indices

Interpreting the overall quality of a partition after having applied a fuzzy clustering algorithm to the symbolic interval data is an important problem in clustering analysis.

3.2.1.1. Overall heterogeneity index. The part of the overall dispersion without clustering (Tl, $l = 1,2$) that corresponds to the dispersion of the partition after clustering (Bl, $l = 1,2$), with each cluster represented by its prototype, is defined as

$$R^l = \frac{B^l}{T^l} = \frac{B^l}{B^l + W^l}\quad (l = 1,2) \tag{20}$$

The fuzzy c-means clustering algorithms are designed so as to maximize R^l $(l = 1,2)$. A greater value of R^l $(l = 1,2)$ signifies more homogeneous clusters and better elements of a cluster P_i represented by its representative $\boldsymbol{g}_i$.

3.2.1.2. Overall heterogeneity indices regarding single variables. The proportion of the overall dispersion without clustering (T_j^l, $l = 1,2$) concerning the jth variable that corresponds to the dispersion of the partition after clustering concerning the jth variable, with each cluster represented by its prototype (B_j^l, $l = 1,2$), is defined as

$$COR^l(j) = \frac{B_j^l}{T_j^l} = \frac{B_j^l}{B_j^l + W_j^l}\quad (l = 1,2) \tag{21}$$

By comparing the value of $COR^l(j)$ $(l = 1,2)$ with the value of the general index R^l $(l = 1,2)$, which measures the average discriminant power of all variables, the discriminant power of the jth variable may be evaluated as being above or below the average.

The relative contribution of the jth variable to the between-cluster fuzzy sum of squares B is given by

$$CTR^l(j) = \frac{B_j^l}{B^l} \quad (l = 1, 2) \tag{22}$$

Notice that $\sum_{j=1}^{p} CTR^l(j) = 1$. A high value of $CTR^l(j)$ $(l = 1, 2)$ indicates that the jth variable provides an important contribution to the separation of the representatives of the clusters.

An interesting case arises when $COR^l(j)$ $(l = 1, 2)$ has a low value and $CTR^l(j)$ $(l = 1, 2)$ is large: this means that the jth variable has a low discriminant power, although it makes an important contribution to the sum of squares (Celeux et al., 1989).

3.2.2. Fuzzy cluster interpretation indices

Another important problem in clustering analysis is evaluating the homogeneity and eccentricity of the individual clusters of a partition after having applied a fuzzy clustering algorithm to the symbolic interval data.

3.2.2.1. Cluster heterogeneity indices. The proportion of the overall fuzzy sum of squares explained by cluster P_i is given by

$$T^l(i) = \frac{T_i^l}{T^l} \quad (l = 1, 2) \tag{23}$$

The relative contribution of a cluster P_i to the between-cluster fuzzy sum of squares is measured by the ratio

$$B^l(i) = \frac{B_i^l}{B^l} \quad (l = 1, 2) \tag{24}$$

A high value of $B^l(i)$ $(l = 1, 2)$ indicates that cluster P_i is quite distant from the global center in comparison to the totality of all clusters.

The relative contribution of cluster P_i to the within-cluster fuzzy sum of squares is given by

$$W^l(i) = \frac{W_i^l}{W^l} \quad (l = 1, 2) \tag{25}$$

A relatively high value of $W^l(i)$ indicates that cluster P_i is relatively heterogeneous in comparison with the other classes.

Notice that $\sum_{i=1}^{c} T^l(i) = \sum_{i=1}^{c} B^l(i) = \sum_{i=1}^{c} J^l(i) = 1$.

3.2.2.2. Cluster heterogeneity regarding single variables. The heterogeneity of clusters may be different for distinct variables. This can be evaluated by considering the previously proposed indices for a single variable j alone. The proportion of the discriminant power of the jth variable with respect to cluster P_i is given by

$$COR^l(j, i) = \frac{B_{ij}^l}{T_j^l} \quad (l = 1, 2) \tag{26}$$

Notice that $\sum_{i=1}^{c} COR^l(j, i) = COR(j)$. A high value of $COR^l(j, i)$ $(l = 1, 2)$ shows that the jth variable has a relatively homogeneous behaviour within the cluster i.

The relative contribution of the jth variable to the heterogeneity in cluster P_i is given by

$$CTR^l(j, i) = \frac{B_{ij}^l}{B_i^l} \quad (l = 1, 2) \tag{27}$$

Finally, we may consider the relative contribution of the jth variable and cluster P_i to the between-cluster fuzzy sum of squares given by

$$CE^l(j, i) = \frac{B_{ij}^l}{B^l} \quad (l = 1, 2) \tag{28}$$

If $CE^l(j, i)$ $(l = 1, 2)$ is close to 1, the jth variable has a large contribution to the eccentricity of cluster P_i.

4. Experimental results

To show the usefulness of these fuzzy clustering methods, two synthetic interval data sets with linearly non-separable clusters of different shapes and sizes have been drawn. Real applications are then considered. Our aim is to achieve a comparison of the dynamic clustering algorithm considering different adaptive distances between vectors of intervals (adaptive Hausdorff distance, see De Carvalho et al. (2006), one component adaptive city-block distance, see Souza and De Carvalho (2004)) and the fuzzy c-means clustering methods presented in this paper.

An external validity index is used to compare the results furnished by these clustering algorithms. For synthetic interval data sets, rectangles are built from three clusters of points drawn from three bi-variate normal distributions. Next, the *a priori* partition of the objects is known. For the symbolic interval data set describing car models, a *a priori* partition into four groups according to a *car category* is defined. Finally, for the city temperature symbolic interval data set describing minimum and maximum temperatures of 37 cities, a *a priori* partition into four (Guru et al., 2004) groups according to the classification given by human observers is defined.

The idea of external validity is simply to compare the a priori partition with the partition obtained from the clustering algorithm. In this paper, we use the corrected Rand (CR) index (Hubert and Arabie, 1985) for comparing two partitions. The CR index measures the similarity between a *a priori* hard partition and a hard partition furnished by a partitioning hard clustering algorithm or obtained from the fuzzy partition furnished by the fuzzy clustering algorithm. CR takes its values on the interval $[-1, 1]$, where the value 1 indicates perfect agreement between partitions, whereas values near 0 (or negatives) correspond to cluster agreement found by chance (Milligan, 1996).

4.1. Synthetic symbolic interval data sets

In this paper, we consider the same data point configurations presented in (Souza and De Carvalho, 2004). Two data sets of 350 points in $\Re^2$ were constructed. In each data set, the 350 points are drawn from three bi-variate normal distributions of independent components. There are three clusters of unequal sizes and shapes: two clusters with an ellipsoidal shape and size 150 and one cluster with a spherical shape and size 50.

Data set 1 shows well-separated clusters (Fig. 1, left side). The data points of each cluster in this data set were drawn according to the following parameters:

(a) Class 1: $\mu_1 = 28$, $\mu_2 = 22$, $\sigma_1^2 = 100$ and $\sigma_2^2 = 9$;
(b) Class 2: $\mu_1 = 60$, $\mu_2 = 30$, $\sigma_1^2 = 9$ and $\sigma_2^2 = 144$;
(c) Class 3: $\mu_1 = 45$, $\mu_2 = 38$, $\sigma_1^2 = 9$ and $\sigma_2^2 = 9$.

Data set 2 shows overlapping clusters (Fig. 1, right side). The data points of each cluster in this data set were drawn according to the following parameters:

(a) Class 1: $\mu_1 = 45$, $\mu_2 = 22$, $\sigma_1^2 = 100$ and $\sigma_2^2 = 9$;
(b) Class 2: $\mu_1 = 60$, $\mu_2 = 30$, $\sigma_1^2 = 9$ and $\sigma_2^2 = 144$;
(c) Class 3: $\mu_1 = 52$, $\mu_2 = 38$, $\sigma_1^2 = 9$ and $\sigma_2^2 = 9$.

In order to build interval data sets from data sets 1 and 2, each point (z_1, z_2) of these data sets is considered as the 'seed' of a rectangle. Each rectangle is therefore a vector of two intervals defined by: $([z_1 - \gamma_1/2, z_1 + \gamma_1/2], [z_2 - \gamma_2/2, z_2 + \gamma_2/2])$.

The parameters γ_1 and γ_2 are the width and the height of the rectangle. They are drawn randomly within a given range of values. Fig. 2 shows two synthetic interval data sets built from data set 1 and data set 2 when γ_1 and γ_2 are drawn randomly from $[1, 8]$.

In the framework of a Monte Carlo experiment, 100 replications of the previous process were carried out for parameters γ_1 and γ_2, drawn randomly 100 times from each of the following intervals: $[1, 8]$, $[1, 16]$, $[1, 24]$, $[1, 32]$, $[1, 40]$. This process has also been repeated for seeds taken from data set 1 and data set 2.

Dynamic hard clustering algorithms considering different adaptive distances (adaptive Hausdorff distance and one component adaptive city-block distance) between vectors of intervals and the (adaptive and non-adaptive) fuzzy clustering methods presented in this paper have been performed on these data sets. The three hard cluster partitions obtained with these clustering methods were compared with the 3-class partition known a priori. The comparison index used is the corrected Rand index *CR*. For each 100 replications, the average corrected Rand index *CR* is calculated.

Table 1 gives the values of the average (and standard-deviation) of the *CR* index obtained with dynamic clustering algorithms considering different adaptive distances and the (adaptive and non-adaptive) fuzzy clustering methods presented in this paper for interval data sets 1 and 2 as well as γ_1 and γ_2 drawn from $[1, 8]$, $[1, 16]$, $[1, 24]$, $[1, 32]$, $[1, 40]$.

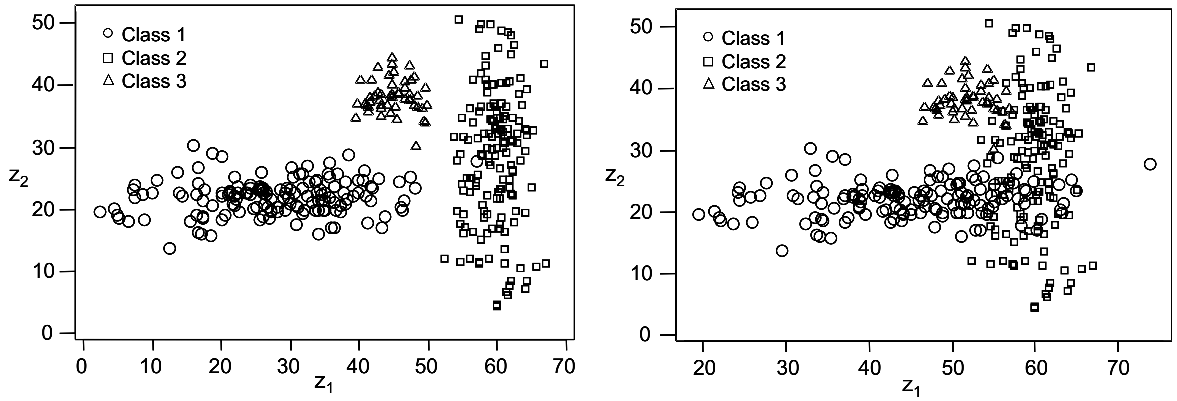

Fig. 1. Seed data sets 1 and 2 showing, respectively, well-separated and overlapping classes.

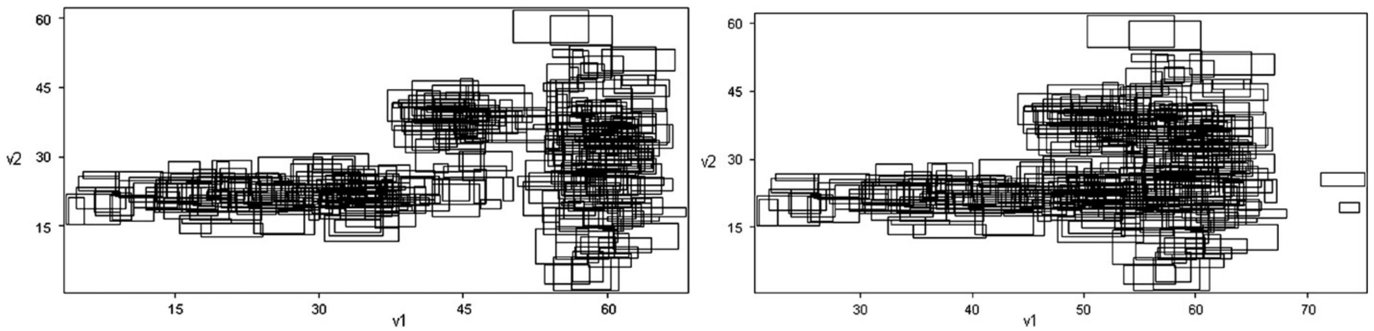

Fig. 2. Interval data sets 1 and 2, showing, respectively, well-separated (left side) and overlapping (right side) classes.

Table 1
Comparison of the methods according to the average (and standard-deviation) of the corrected Rand index

Predefined intervals	Interval data set 1				Interval data set 2			
	Hard clustering methods		Fuzzy *c*-means clustering methods		Hard clustering methods		Fuzzy *c*-means clustering methods	
	L_1	Hausd.	IFCM	IFCMADC	L_1	Hausd.	IFCM	IFCMADC
[1, 8]	0.935 (0.0005)	0.923 (0.0010)	0.837 (0.1331)	0.964 (0.1197)	0.470 (0.0050)	0.448 (0.0019)	0.430 (0.0844)	0.832 (0.1491)
[1, 16]	0.931 (0.0009)	0.931 (0.0007)	0.844 (0.0443)	0.979 (0.0119)	0.432 (0.0014)	0.434 (0.0023)	0.443 (0.0378)	0.812 (0.0361)
[1, 24]	0.892 (0.0058)	0.909 (0.0011)	0.822 (0.0481)	0.957 (0.0206)	0.406 (0.0015)	0.418 (0.0019)	0.425 (0.0322)	0.739 (0.0441)
[1, 32]	0.773 (0.0114)	0.912 (0.0011)	0.761 (0.0529)	0.919 (0.0244)	0.389 (0.0013)	0.412 (0.0017)	0.409 (0.0315)	0.630 (0.0536)
[1, 40]	0.701 (0.0073)	0.886 (0.0032)	0.739 (0.0479)	0.868 (0.0323)	0.373 (0.0024)	0.393 (0.0024)	0.390 (0.0339)	0.527 (0.0704)

Regarding the data configurations presenting well-separated classes, in each case the average *CR* indices are better with adaptive distances. Moreover, the IFCMADC clustering algorithm shows better *CR* indices than the dynamic hard clustering algorithms (with adaptive Hausdorff or city-block distances) regardless of the range of the predefined intervals in Table 1.

Notice that, for data configurations presenting overlapping classes, the IFCMADC clustering algorithm clearly outperforms the other methods and, in this case, the IFCM clustering method presents almost the same performance as the dynamic hard clustering methods based on adaptive (Hausdorff or city-block) distances.

4.2. Symbolic interval data sets

For the purpose of validating the proposed methods for efficiency, we have conducted several experiments on the following data sets of type interval: car symbolic interval data set and city temperature symbolic interval data set.

4.2.1. Car symbolic interval data set

The car symbolic interval data set consists of a set of 33 car models described by 8 interval variables, 2 categorical multi-valued variables and one nominal variable (De Carvalho et al., 2006). In this application, the 8 interval variables – *Price, Engine Capacity, Top Speed, Acceleration, Step, Length, Width* and *Height* – were considered for clustering purposes, the nominal variable *Car Category* was used as a *a priori* classification.

Dynamic hard clustering algorithms considering different adaptive distances (one component adaptive city-block distance and adaptive Husdorff distance) between vectors of intervals as well as the IFCMADC clustering algorithm were performed on this data set. The 4-cluster partitions obtained with these clustering methods were compared with the 4-cluster partition known *a priori*. The comparison index used is the corrected Rand index *CR*. 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/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.
`Sports`: 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.

Each clustering method is run (until the convergence to a stationary value of the adequacy criterion) 100 times and the best result according to the corresponding adequacy criterion is selected. For the IFCMADC clustering algorithm, the parameter m was set to 2. The corrected Rand index CR is calculated for the best result.

Table 2 shows the clusters (individual labels) given by the adaptive (one component L_1, Hausdorff and IFCMADC) methods.

The CR indices obtained from the results displayed in Table 2 are 0.56, 0.56 and 0.52 for the adaptive one component L_1, Hausdorff and IFCMADC methods, respectively. Notice that, for this data set, the dynamic hard clustering algorithms with adaptive (one component city-block and Hausdorff) distances slightly outperform the IFCMADC algorithm. Moreover, the same partition is furnished by the dynamic hard clustering algorithms (either with the adaptive Hausdorff distance or with the one component adaptive city-block distance).

4.2.2. City temperature symbolic interval data set

The city temperature symbolic interval data set (Guru et al., 2004) gives the minimum and the maximum monthly temperatures of 37 cities in degrees centigrade. A *a priori* classification given by a panel of human observers is as follows:

Class 1: *Bahraim Bombay Cairo Calcutta Colombo Dubai Hong Kong Kula Lampur Madras Manila Mexico Nairobi New Delhi Sydney*

Class 2: *Amsterdam Athens Copenhagen Frankfurt Geneva Lisbon London Madrid Moscow Munich New York Paris Rome San Francisco Seoul Stockholm Tokyo Toronto Vienna Zurich*

Class 3: *Mauritius*

Class 4: *Tehran*

Cities belonging to Class 1 are mainly located between 0° and 40° latitudes and the cities that are classified as Class 2 are mainly located between 40° and 60° latitudes. Some cities, which are closer to the sea coast and are located between latitudes 0° and 40°, are classified as members of Class 2. Mauritius island and Tehran are classified as members of singleton Classes 3 and 4, respectively. The clusters obtained using the approach proposed by Guru et al. (2004) are in complete accordance with these *a priori* classes provided by the panel of human observers.

With this city temperature symbolic interval data set, each fuzzy c-means clustering method was run until the convergence to a stationary value of the criterion W^l $(l = 1, 2)$ 60 times and the best result according to the corresponding adequacy criterion was selected. The parameter m was set to 2.

Table 2
Clustering results for the car symbolic interval data set

Method	Cluster 1	Cluster 2	Cluster 3	Cluster 4
IFCMADC	4/S 11/S 15/S 16/S 19/S 27/S	1/U 12/U 13/U 14/B 17/U 24/U 25/U 28/U 29/U 31/U	6/B 7/L 9/L 10/L 20/S 22/L 23/L	2/B 3/L 5/U 8/B 18/L 21/B 26/B 30/B 32/B 33/L
L_1	12/U 13/U 17/U 24/U 25/U 28/U 29/U 31/U	1/U 2/B 3/L 5/U 8/B 14/B 18/L 21/B 26/B 30/B 32/B 33/L	6/B 7/L 9/L 10/L 22/L 23/L	4/S 11/S 15/S 16/S 19/S 20/S 27/S
Hausdorf	1/U 2/B 3/L 5/U 8/B 14/B 18/L 21/B 26/B 30/B 32/B 33/L	12/U 13/U 17/U 24/U 25/U 28/U 29/U 31/U	4/S 11/S 15/S 16/S 19/S 20/S 27/S	6/B 7/L 9/L 10/L 22/L 23/L

Table 3
Hard partition in 4 clusters obtained from the fuzzy c-means clustering algorithms

Partition	IFCM	IFCMADC
Cluster 1	Bahrain, Cairo, Hong Kong, Mexico, Nairobi, New Delhi, Sydney	Bahrain, Bombay, Calcutta, Colombo, Dubai, Hong Kong, Kula Lampur, Madras, Manila, New Delhi
Cluster 2	Amsterdam, Copenhagen, Frankfurt, Geneva, London, Moscow, Munich, New York, Paris, Stockholm, Toronto, Vienna	Amsterdam, Copenhagen, Frankfurt, Geneva, London, Moscow, Munich, Paris, Stockholm, Toronto, Vienna
Cluster 3	Athens, Lisbon, Madrid, Rome, San Francisco, Seoul, Tehran, Tokyo, Zurich	Cairo, Mauritius, Mexico, Nairobi, Sydney
Cluster 4	Bombay, Calcutta, Colombo, Dubai, Kula Lampur, Madras, Manila, Mauritius	Athens, Lisbon, Madrid, New York, Rome, San Francisco, Seoul, Tehran, Tokyo, Zurich

Please cite this article in press as: de Carvalho, F.d.A.T., Fuzzy c-means clustering methods for symbolic interval data, Pattern Recog. Lett. (2006), doi:10.1016/j.patrec.2006.08.014

Table 3 shows the hard partitions in 4 clusters obtained from the fuzzy partitions furnished by the (non-adaptive and adaptive) fuzzy *c*-means clustering algorithms.

The *CR* indices obtained from the results displayed in Table 3 are 0.46 and 0.50 for the IFCM and IFCMADC methods, respectively. In the partition furnished by the IFCM clustering algorithm, Cluster 1 and Cluster 4 are a splitting of *a priori* Class 1, whereas Cluster 2 and Cluster 3 are a splitting of *a priori* Class 2. Notice that the IFCM clustering algorithm provides a coherent splitting of *a priori* Classes 1 and 2. Cluster 4 includes Mauritius island (*a priori* Class 3) and Cluster 3 includes Teheran (*a priori* Class 4). Cluster 2 is mainly formed by north and central European cities and half of the cities belonging to Cluster 3 come from southern Europe.

Concerning the partition obtained using the IFCMADC clustering algorithm, Cluster 1 and Cluster 3 are also a splitting of *a priori* Class 1, whereas Cluster 2 and Cluster 4 are a splitting of *a priori* Class 2. Notice that the IFCMADC clustering algorithm also provides a coherent splitting of *a priori* Classes 1 and 2. Cluster 3 includes Mauritius island (*a priori* Class 3) and Cluster 4 includes Teheran (*a priori* Class 4). Again, Cluster 2 is mainly formed by north and central European cities and Cluster 4 has many cities from southern Europe.

4.3. Fuzzy partition and cluster interpretation: the city temperature symbolic interval data set

The merit of the interpretation indices presented in this paper will be highlighted here through the results obtained with application of the adaptive and non-adaptive fuzzy *c*-means clustering methods to the city temperature symbolic interval data set to obtain a fuzzy partition in 4 clusters.

Table 4
Overall heterogeneity index for the fuzzy *c*-means methods

Method	Non-adaptive ($l = 1$)	Adaptive ($l = 1$)
R^l ($l = 1, 2$)	0.821249	0.808690

4.3.1. Fuzzy partition interpretation

Table 4 shows that the proportion of the overall fuzzy SSQ explained by the partition for the adaptive and non-adaptive fuzzy *c*-means clustering methods. As the values of the overall heterogeneity index are close to 1, the individuals belonging to the clusters are well represented by the corresponding cluster prototypes in both cases.

Comparing the values of $COR^1(j)$ with the value of R^1 (see Tables 4 and 5) for the partition obtained with the IFCM clustering method, we may conclude that the discriminant power of the symbolic interval variables 5 (*May*), 6 (*June*), 7 (*July*), 8 (*August*) and 9 (*September*) are below the average discriminate power of all variables. All the other variables have a discriminant power above the average. From Table 5, we can see that variables 1 (*January*), 2 (*February*), 3 (*March*), 11 (*November*) and 12 (*December*) provide the greatest contribution to the separation of the prototypes of the clusters.

Comparing the values of $COR^2(j)$ with the values of R^2 (see Table 4) for the partitions obtained using the IFCMADC clustering method (see Table 6), we may conclude that the discriminant power of the symbolic interval variables 3 (*March*), *4* (*April*), *10* (*October*), 11 (*November*) and 12 (*December*) is above the average. Interval variables 1 (*January*) and 2 (*February*) have a discriminant power slightly below the average, whereas all the other interval variables have a discriminant power clearly below the average. From Table 6, we can see that variables 4 (*April*), 10 (*October*) and 11 (*November*) provide important contributions to the separation of the prototypes of the clusters.

4.3.2. Cluster interpretation

From Table 7, we can see that in the partition obtained with the IFCM clustering method, the Cluster 1 mean vector is the closest to the global mean vector, while the Cluster 2 mean vector is the farthest away. Moreover, Cluster 2 is the least homogeneous and Cluster 4 is the most homogeneous of the four clusters.

The values in Table 8 show that in the partition obtained from the IFCMADC clustering method, the Cluster 4 mean vector is the closest to the global mean, while the Cluster 2 mean vector is the farthest away. Moreover, Clus-

Table 5
Overall heterogeneity indices concerning the symbolic interval variables for the IFCM clustering method (%)

Variables	1	2	3	4	5	6	7	8	9	10	11	12
$COR^1(j)$	85.5	84.0	86.8	87.8	81.2	72.5	59.7	64.3	64.5	87.5	88.7	85.7
$CTR^1(j)$	13.4	13.5	11.6	8.6	7.0	4.2	2.8	3.5	4.3	8.1	10.2	12.6

Table 6
Overall heterogeneity indices concerning the symbolic interval variables for the IFCMADC clustering method (%)

Variables	1	2	3	4	5	6	7	8	9	10	11	12
$COR^2(j)$	80.2	80.7	83.3	85.9	78.8	78.3	67.7	75.0	69.4	87.3	84.1	81.7
$CTR^2(j)$	8.0	8.3	9.8	12.0	7.3	7.1	4.1	5.9	4.4	13.6	10.4	8.8

Please cite this article in press as: de Carvalho, F.d.A.T., Fuzzy *c*-means clustering methods for symbolic interval data, Pattern Recog. Lett. (2006), doi:10.1016/j.patrec.2006.08.014

Table 7
Cluster heterogeneity indices for the IFCM clustering method

Cluster	Cardinal	$T^1(i)$	$B^1(i)$	$W^1(i)$
1	7	0.045920	0.005461	0.231807
2	12	0.617941	0.686965	0.300817
3	9	0.150933	0.124323	0.273185
4	9	0.185206	0.183250	0.194190

Table 8
Cluster heterogeneity indices for the IFCMADC clustering method

Cluster	Cardinal	$T^2(i)$	$B^2(i)$	$W^2(i)$
1	11	0.385323	0.419173	0.242237
2	11	0.435040	0.470057	0.287018
3	5	0.080593	0.052354	0.199962
4	10	0.099043	0.058415	0.270782

ter 3 is the most homogeneous and Cluster 2 is the least homogeneous of the four clusters.

Table 9 shows the cluster heterogeneity indices concerning the variables for the IFCM clustering method.

From this table, variable 2 (*February*) presents a homogeneous behaviour within Cluster 1 and provides the greatest contribution to the separation of the Cluster 1 mean vector from the global mean vector.

All the variables presented a homogeneous behaviour within Cluster 2, while interval variables 1 (*January*), 2 (*February*), 3 (*March*), 11 (*November*) and 12 (*December*) play the most important role in the heterogeneity of this cluster, as well as in the separation of Cluster 2 mean vector from the global mean vector.

Concerning Cluster 3, interval variables 1 (*January*), 3 (*March*), 11 (*November*) and 12 (*December*) play the most important role in the heterogeneity of this cluster.

Finally, interval variables 1 (*January*), 2 (*February*) and 3 (*March*) present the most homogeneous behaviour within Cluster 4 and, together with interval variable 12 (*December*), play the most important role in the heterogeneity of this cluster, as well as in the separation of Cluster 4 mean vector from the global mean vector.

Table 10 shows the cluster heterogeneity indices concerning the variables for the IFCMADC clustering method.

From this table, we can see that all the interval variables presented a quite homogeneous behaviour within Cluster 1, except interval variable 8 (*August*). Interval variables 2 (*February*), 3 (*March*), 4 (*April*), 10 (*October*) and 11 (*November*) play the most important role in the heterogeneity of this cluster, as well as in the separation of Cluster 1 mean vector from the global mean vector.

Table 9
Cluster heterogeneity indices concerning the variables for the IFCM method (%)

	Cluster 1			Cluster 2			Cluster 3			Cluster 4		
	COR	CTR	CE	COR	CTR	CE	COR	CTR	CE	COR	CTR	CE
1	0.46	13.1	0.07	52.6	12.0	8.2	9.2	11.6	1.4	23.3	19.9	3.6
2	1.0	30.5	0.17	50.6	11.9	8.2	6.8	8.8	1.1	25.5	22.4	4.1
3	0.5	12.4	0.07	53.9	10.5	7.2	11.3	12.1	1.5	21.1	15.4	2.8
4	0.72	12.9	0.07	55.6	7.9	5.4	12.5	9.9	1.2	18.9	10.1	1.8
5	0.33	05.3	0.03	59.4	7.5	5.1	13.7	9.5	1.2	7.8	3.7	0.7
6	0.19	2.0	0.01	57.5	4.8	3.3	11.5	5.3	0.7	3.2	1.0	0.2
7	0.17	1.4	0.01	52.9	3.6	2.4	5.9	2.2	0.3	0.7	0.2	0.03
8	0.26	2.6	0.01	57.0	4.6	3.1	6.6	2.9	0.4	0.4	0.1	0.02
9	0.10	1.2	0.01	56.2	5.4	3.7	5.6	2.9	0.4	2.6	0.9	0.2
10	0.26	4.4	0.02	68.5	9.2	6.3	12.2	9.0	1.1	6.5	3.2	0.6
11	0.28	6.0	0.03	64.3	10.8	7.4	11.7	10.9	1.3	12.4	7.8	1.4
12	0.29	7.9	0.04	54.6	11.7	8.0	12.3	14.6	1.8	18.5	14.9	2.7

Table 10
Cluster heterogeneity indices concerning the variables for the IFCMADC clustering method (%)

	Cluster 1			Cluster 2			Cluster 3			Cluster 4		
	COR	CTR	CE	COR	CTR	CE	COR	CTR	CE	COR	CTR	CE
1	30.0	7.1	3.0	36.7	7.8	3.7	6.8	13.0	0.7	6.7	11.3	0.7
2	41.9	10.2	4.3	21.6	4.7	2.2	13.9	27.3	1.4	3.2	5.6	0.3
3	36.4	10.3	4.3	31.3	7.9	3.7	7.4	16.8	0.9	8.2	16.6	1.0
4	39.4	13.1	5.5	36.3	10.8	5.1	3.1	8.2	0.4	7.1	16.9	1.0
5	37.8	8.4	3.5	34.7	6.9	3.2	0.04	0.07	0.0	6.3	10.0	0.6
6	28.4	6.1	2.6	46.8	9.0	4.2	1.3	2.3	0.1	1.7	2.6	0.1
7	22.8	3.3	1.4	41.3	5.4	2.5	2.6	3.0	0.1	0.9	1.0	0.06
8	12.1	2.3	0.9	59.3	9.9	4.6	3.3	5.0	0.2	0.2	0.3	0.02
9	44.8	6.9	2.9	23.6	3.2	1.5	0.9	1.1	0.0	0.1	0.2	0.01
10	40.8	15.2	6.4	43.3	14.3	6.7	0.2	0.8	0.0	2.9	7.9	0.4
11	35.4	10.5	4.4	37.3	9.8	4.6	5.0	11.9	0.6	6.3	13.4	0.8
12	25.0	6.4	2.7	44.1	10.1	4.7	4.9	10.2	0.5	7.6	13.9	0.8

In Cluster 2, interval variables 2 (*February*) and 9 (*September*) presented the least homogeneous behaviour, while interval variables 6 (*June*), 7 (*July*), 8 (*August*), 10 (*October*) and 12 (*December*) presented the most homogeneous behaviour. Interval variables 4 (*April*), 10 (*October*), 11 (*November*) and 12 (*December*) play the most important role in the heterogeneity of the Cluster 2, as well as in the separation of Cluster 2 mean vector from the global mean vector.

Moreover, interval variable 2 (*February*) plays the most important role in the heterogeneity of Cluster 3. Finally, interval variables 3 (*March*), 4 (*April*), 11 (*November*) and 12 (*December*) play the most important role in the heterogeneity of the Cluster 4.

5. Concluding remarks

The main contributions of this paper are the introduction of adaptive and non-adaptive fuzzy *c*-means clustering algorithms (IFCM and IFCMADC) for symbolic interval data and various fuzzy partition and cluster interpretation tools that are suitable for these fuzzy clustering algorithms. The IFCM clustering algorithm starts from an initial fuzzy partition and alternates a representation step, where the partition is fixed and the algorithm gives the solution for the best prototype of each class, and an allocation step, where the prototypes of the classes are fixed and the algorithm provides the solution for the best membership degree of each pattern in each class, until convergence when the adequacy criterion reaches a stationary value representing a local minimum.

The IFCMADC clustering method is based on a suitable adaptive squared Euclidean distance for each class. The main idea of this method is that there is a different distance associated to each cluster for comparing clusters and their prototypes. This distance changes at each iteration. The method starts from an initial partition and alternates a representation step and an allocation step until convergence, when the adequacy criterion reaches a stationary value representing a local minimum. The representation step has now two stages. In the first stage, the membership degree of each pattern in each class and the distances are fixed and the algorithm gives the solution for the best prototype of each class. In the second stage, the membership degree of each pattern in each class and the prototypes of the classes are fixed and the algorithm provides the solution for the best distance of each class. Finally, in the allocation step, the prototypes and the distances are fixed and the algorithm gives the solution for the best membership degree of each pattern in each class.

The problem of interpreting and evaluating the obtained fuzzy partition has been addressed. We were able to define the overall fuzzy sum of squares and the fuzzy sum of squares within and between clusters for symbolic interval data and to show that the overall fuzzy sum of squares decomposes into a fuzzy sum of squares within a cluster plus a fuzzy sum of squares between clusters. Based on this decomposition, a family of fuzzy partition and cluster interpretation indices for the adaptive and non-adaptive fuzzy *c*-means clustering methods for symbolic interval data have been introduced. These new indices constitute a suitable adaptation of the indices introduced for interpreting and evaluating partitions furnished by the standard hard *c*-means clustering method.

Experiments with real and synthetic symbolic interval data sets showed the usefulness of this clustering method. The accuracy of the results furnished by these fuzzy *c*-means clustering algorithms is assessed by the *CR* index. Concerning the synthetic interval data sets, the *CR* index is calculated for these fuzzy *c*-means clustering algorithms and compared with the results provided by the dynamic hard clustering algorithms considering different adaptive distances (adaptive Hausdorff distance and one component adaptive city-block distance) in the framework of a Monte Carlo simulation with 60 replications. Symbolic interval data configurations showing well-separated and overlapping classes are considered. Concerning the data configurations presenting well-separated classes, in each case the average *CR* indices are better with adaptive distances. Moreover, the IFCMADC clustering algorithm shows better *CR* indices than the dynamic hard clustering algorithms, regardless of the adaptive (Hausdorff or city-block) distance considered and the range of the predefined intervals. For data configurations presenting overlapping classes, the IFCMADC clustering algorithm clearly outperforms the other methods and, in this case, the IFCM clustering method presents nearly the same performance as the dynamic hard clustering methods based on adaptive (Hausdorff or city-block) distances. Concerning the car interval data set, the dynamic hard clustering methods based on adaptive (Hausdorff or city-block) distances slightly outperform the IFCMADC method. Concerning the city temperature symbolic interval data set with 4 *a priori* classes, the fuzzy *c*-means clustering algorithms provided a coherent splitting of *a priori* Classes 1 and 2. Finally, the application of the adaptive and non-adaptive fuzzy *c*-means clustering methods to the city temperature symbolic interval data set showed the merit of the fuzzy partition and cluster interpretation indices proposed in this paper.

Acknowledgements

The author would like to thank CNPq (Brazilian Agency) for its financial support and the anonymous referees for their helpful comments and suggestions to improve the paper.

Appendix A. Proof of Proposition 2.4

The vectors of weights $\boldsymbol{\lambda}_i = (\lambda_i^1, \ldots, \lambda_i^p)(i = 1, \ldots, c)$, which minimize the clustering criterion W^2 under $\lambda_i^j > 0$ and $\prod_{j=1}^{p} \lambda_i^j = 1$, are updated according to the following expression:

$$\lambda_i^j = \frac{\left\{\prod_{h=1}^p \left[\sum_{k=1}^n (u_{ik})^m ((a_k^h - \alpha_i^h)^2 + (b_k^h - \beta_i^h)^2)\right]\right\}^{\frac{1}{p}}}{\sum_{k=1}^n (u_{ik})^m [(a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2]},$$
$$j = 1, \ldots, p$$

Proof. As the membership degree u_{ik} of each pattern k in cluster P_i, the parameter m and the prototypes $\boldsymbol{g}_i$ of class $P_i\,(i = 1, \ldots, c)$ are fixed, we can rewrite the criterion W^2 as

$$W^2(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p) = \sum_{i=1}^c W_i^2(\boldsymbol{\lambda}_i) \quad \text{with}$$
$$W_i^2(\boldsymbol{\lambda}_i) = W_i^2(\lambda_i^1, \ldots, \lambda_i^p) = \sum_{j=1}^p \lambda_i^j W_{ij}^2$$
$$\text{where } W_{ij}^2 = \sum_{k=1}^n (u_{ik})^m [(a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2].$$

The criterion W^2 being additive, the problem becomes minimizing $W_i^2 (i = 1, \ldots, c)$. Let $g(\lambda_i^1, \ldots, \lambda_i^p) = \prod_{j=1}^p \lambda_i^j - 1 = \lambda_i^1 \times \ldots \times \lambda_i^p - 1$. We want to determine the extremes of $W_i^2(\lambda_i^1, \ldots, \lambda_i^p)$ with the restriction $g(\lambda_i^1, \ldots, \lambda_i^p) = 0$. To do so, we shall use the method of Lagrange multipliers. After some algebra, we conclude that an extreme value of W_i^2 is reached when

$$\lambda_i^j = \frac{\left\{\prod_{h=1}^p W_{ih}^2\right\}^{\frac{1}{p}}}{W_{ij}^2}$$
$$= \frac{\left\{\prod_{h=1}^p \left[\sum_{k=1}^n (u_{ik})^m [(a_k^h - \alpha_i^h)^2 + (b_k^h - \beta_i^h)^2]\right]\right\}^{\frac{1}{p}}}{\sum_{k=1}^n (u_{ik})^m [(a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2]}$$
$$(j = 1, \ldots, p)$$

This extreme value is $W_i^2(\lambda_i^1, \ldots, \lambda_i^p) = \sum_{j=1}^p \lambda_i^j W_{ij}^2 = p\{W_{i1}^2 \times \ldots \times W_{ip}^2\}^{\frac{1}{p}}$.

As $W_i^2(1, \ldots, 1) = \sum_{j=1}^p W_{ij}^2 = W_{i1}^2 + \cdots + W_{ip}^2$, and as it is well known that the arithmetic mean is greater than the geometric mean, i.e., $\frac{1}{p}(W_{i1}^2 + \cdots + W_{ip}^2) > \{W_{i1}^2 \ldots W_{ip}^2\}^{\frac{1}{p}}$ (the equality holds only if $W_{i1}^2 = \ldots = W_{ip}^2$), we conclude that this extreme is a minimum.

Appendix B. Proof of Proposition 3.1

For $l = 1, 2$, the following relations hold for all j and i:

$$T^l = W^l + B^l, \quad T_i^l = B_i^l + W_i^l, \quad T_j^l = B_j^l + W_j^l,$$
$$T_{ij}^l = B_{ij}^l + W_{ij}^l$$

Proof. We will start showing that $T^1 = W^1 + B^1$ holds. We have,

$$\begin{aligned}(a_k^j - \alpha^j)^2 + (b_k^j - \beta^j)^2 = {} & [(a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2] \\ & + [(\alpha_i^j - \alpha^j)^2 + (\beta_i^j - \beta^j)^2] \\ & + 2[(a_k^j - \alpha_i^j)(\alpha_i^j - \alpha^j)] \\ & + 2[(b_k^j - \beta_i^j)(\beta_i^j - \beta^j)]\end{aligned}$$

Then, from Eq. (8),

$$\begin{aligned}T^1 &= \sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m \sum_{j=1}^p [(a_k^j - \alpha^j)^2 + (b_k^j - \beta^j)^2] \\ &= W^1 + B^1 + 2\sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m \sum_{j=1}^p [(a_k^j - \alpha_i^j)(\alpha_i^j - \alpha^j)] \\ &\quad + 2\sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m \sum_{j=1}^p [(b_k^j - \beta_i^j)(\beta_i^j - \beta^j)]\end{aligned}$$

We have also,

$$(a_k^j - \alpha_i^j)(\alpha_i^j - \alpha^j) = \alpha_i^j(a_k^j - \alpha_i^j) - \alpha^j(a_k^j - \alpha_i^j)$$
$$\text{and} \quad (b_k^j - \beta_i^j)(\beta_i^j - \beta^j) = \beta_i^j(b_k^j - \beta_i^j) - \beta^j(b_k^j - \beta_i^j)$$

Then,

$$\begin{aligned}&\sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m \sum_{j=1}^p ((a_k^j - \alpha_i^j)(\alpha_i^j - \alpha^j)) \\ &\quad = \sum_{j=1}^p \sum_{i=1}^c \left\{ \alpha_i^j \left[\sum_{k=1}^n (u_{ik})^m a_k^j - \alpha_i^j \sum_{k=1}^n (u_{ik})^m\right] \right. \\ &\qquad \left. - \alpha^j \left[\sum_{k=1}^n (u_{ik})^m a_k^j - \alpha_i^j \sum_{k=1}^n (u_{ik})^m\right]\right\}\end{aligned}$$

and

$$\begin{aligned}&\sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m \sum_{j=1}^p ((b_k^j - \beta_i^j)(\beta_i^j - \beta^j)) \\ &\quad = \sum_{j=1}^p \sum_{i=1}^c \left\{ \beta_i^j \left[\sum_{k=1}^n (u_{ik})^m b_k^j - \beta_i^j \sum_{k=1}^n (u_{ik})^m\right] \right. \\ &\qquad \left. - \beta^j \left[\sum_{k=1}^n (u_{ik})^m b_k^j - \beta_i^j \sum_{k=1}^n (u_{ik})^m\right]\right\}\end{aligned}$$

As $\alpha_i^j \sum_{k=1}^n (u_{ik})^m = \sum_{k=1}^n (u_{ik})^m a_k^j$ and $\beta_i^j \sum_{k=1}^n (u_{ik})^m = \sum_{k=1}^n (u_{ik})^m b_k^j$, it follows that $\sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m \sum_{j=1}^p (a_k^j - \alpha_i^j)(\alpha_i^j - \alpha^j) = 0$ and $\sum_{i=1}^c \sum_{k=1}^n (u_{ik})^m \sum_{j=1}^p (b_k^j - \beta_i^j)(\beta_i^j - \beta^j) = 0$, and then $T^1 = W^1 + B^1$.

The other expressions can be easily obtained in a similar way.

References

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

Billard, L., Diday, E., 2003. From the statistics of data to the statistics of knowledge: Symbolic data analysis. J. Amer. Statist. Assoc. 98 (462), 470–487.

Bock, H.-H., 2002. Clustering algorithms and Kohonen maps for symbolic data. J. Japanese Soc. Comput. Statist. 15, 1–13.

Bock, H.H., Diday, E., 2000. Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data. Springer-Verlag, Heidelberg.

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

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

Chavent, M., Lechevallier, Y., 2002. Dynamical clustering algorithm of interval data: Optimization of an adequacy criterion based on Hausdorff distance. In: Sokolowski, A., Bock, H.-H. (Eds.), Classification, Clustering and Data Analysis. Springer, Heidelberg, pp. 53–59.

Please cite this article in press as: de Carvalho, F.d.A.T., Fuzzy *c*-means clustering methods for symbolic interval data, Pattern Recog. Lett. (2006), doi:10.1016/j.patrec.2006.08.014

De Carvalho, F.A.T., Souza, R.M.C.R., Chavent, M., Lechevallier, Y., 2006. Adaptive Hausdorff distances and dynamic clustering of symbolic data. Pattern Recognition Lett. 27 (3), 167–179.

Diday, E., Govaert, G., 1977. Classification automatique avec distances adaptatives. R.A.I.R.O. Inform. Comput. Sci. 11 (4), 329–349.

Diday, E., Brito, P., 1989. Symbolic cluster analysis. In: Opitz, O. (Ed.), Conceptual and Numerical Analysis of Data. Springer-Verlag, Heidelberg, pp. 5–84.

Dunn, J.C., 1974. A fuzzy relative to the ISODATA process and its use in detecting compact, well-separated clusters. J. Cybernet. 3, 32–57.

El-Sonbaty, Y., Ismail, M.A., 1998. Fuzzy clustering for symbolic data. IEEE Trans. Fuzzy Systems 6, 195–204.

Gordon, A.D., 1999. Classification. Chapman and Hall/CRC, Boca Raton, FL.

Gordon, A.D., 2000. An iteractive relocation algorithm for classifying symbolic data. In: Gaul, W. et al. (Eds.), Data Analysis: Scientific Modeling and Practical Application. Springer-Verlag, Berlin, pp. 7–23.

Gowda, K.C., Diday, E., 1991. Symbolic clustering using a new dissimilarity measure. Pattern Recognition 24 (6), 567–578.

Gowda, K.C., Diday, E., 1992. Symbolic clustering using a new similarity measure. IEEE Trans. Systems Man Cybernet. 22, 368–378.

Gowda, K.C., Ravi, T.R., 1995a. Divisive clustering of symbolic objects using the concepts of both similarity and dissimilarity. Pattern Recognition 28 (8), 1277–1282.

Gowda, K.C., Ravi, T.R., 1995b. Agglomerative clustering of symbolic objects using the concepts of both similarity and dissimilarity. Pattern Recognition Lett. 16, 647–652.

Gowda, K.C., Ravi, T.R., 1999a. Clustering of symbolic objects using gravitational approach. IEEE Trans. Systems Man Cybernet. 29 (6), 888–894.

Guru, D.S., Kiranagi, B.B., 2005. Multivalued type dissimilarity measure and concept of mutual dissimilarity value for clustering symbolic patterns. Pattern Recognition 38, 151–256.

Guru, D.S., Kiranagi, B.B., Nagabhushan, P., 2004. Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns. Pattern Recognition Lett. 25, 1203–1213.

Gustafson, D.E., Kessel, W.C., 1979. Fuzzy clustering with a fuzzy covariance matrix. In: Proc. IEEE Conf. Decision Contr., San Diego, CA, 761–766.

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

Ichino, M., Yaguchi, H., 1994. Generalized Minkowski metrics for mixed feature type data analysis. IEEE Trans. Systems Man Cybernet. 24 (4), 698–708.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: A review. ACM Comput. Surveys 31 (3), 264–323.

Milligan, G.W., 1996. Clustering validation: Results and implications for applied analysis. In: Arabie, P., Hubert, L.J., De Soete, G. (Eds.), Clustering and Classification. Word Scientific, Singapore, pp. 341–375.

Ralambondrainy, H., 1995. A conceptual version of the *k*-means algorithm. Pattern Recognition Lett. 16, 1147–1157.

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

Verde, R., De Carvalho, F.A.T., Lechevallier, Y., 2001. A dynamical clustering algorithm for symbolic data. In: Tutorial on Symbolic Data Analysis held during the 25th Annual Conference of the Gesellschaft für Klassifikation, University of Munich, March 13, 2001.

Yang, M.-S., Hwang, P.-Y., Chen, D.-H., 2004. Fuzzy clustering algorithms for mixed feature variables. Fuzzy Sets Systems 141, 301–317.