A robust fuzzy *k*-means clustering model for interval valued data

Pierpaolo D'Urso[1] and Paolo Giordani[2]

[1] Dipartimento di Scienze Economiche, Gestionali e Sociali, Università degli Studi del Molise, Via De Sanctis, 86100 Campobasso, Italy
[2] Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università di Roma "La Sapienza", P.le Aldo Moro, 5, 00185 Rome, Italy

Summary

In this paper a robust fuzzy *k*-means clustering model for interval valued data is introduced. The peculiarity of the proposed model is the capability to manage anomalous interval valued data by reducing the effect of such outliers in the clustering model. In the interval case, the concept of anomalous data involves both the center and the width (the radius) of an interval. In order to show how our model works, the results of some applications to synthetic and real interval valued data are discussed.

Keywords: Fuzzy *k*-means, Robust clustering, Interval valued data, Noise center, Noise radius, Noise cluster

1 Introduction

In the recent literature, several references on statistical analysis of data with complex structure (complex data) (data of possibly different nature, i.e. interval valued data, symbolic data, fuzzy data, mixed feature data) may be found (see, for instance, Bock & Diday, 2000). In particular, focusing on the principal multivariate exploratory statistical analyses, we have the following references:

- Cluster Analysis (Auephanwiriyakul & Keller, 2002; Chavent, 2000; Chavent & Lechevallier, 2002; Coppi & D'Urso, 2002, 2003b; D'Urso & Giordani, 2005b; de Carvalho, 1994; de Carvalho et al., 2004; de Souza & de Carvalho, 2004; Diday, 1988; El-Sonbaty & Ismail, 1998a,b; Gowda & Diday, 1991, 1992; Gowda & Ravi, 1995a,b; 1999a,b; Guru et al., 2004; Hathaway et al., 1996; Hwang, 1989; Ichino & Yaguchi, 1994; Mali & Mitra, 2003; Masson & Denœux, 2004; Pedrycz et al., 1998; Yang & Ko, 1996; Yang & Liu, 1999; Yang et al., 2004;).
- Principal Component Methods (see, e.g., Cazes et al., 1997; Coppi et al., 2005; D'Urso & Giordani, 2004, 2005a; Denœux & Masson, 2004; Giordani & Kiers, 2004; Lauro & Palumbo, 2000, 2005; Lauro et al., 2000; Palumbo & Lauro, 2003, Watada & Yabuuchi, 1997).
- Multidimensional Scaling (see, e.g., Denœux & Masson, 2000; Masson & Denœux, 2002).
- Regression Analysis (see, e.g., Billard & Diday, 2000; Coppi & D'Urso, 2003a; D'Urso & Gastaldi, 2000; 2002; D'Urso, 2003; D'Urso & Giordani, 2003; Hong & Hwang, 2005; Körner & Näther, 1998; Yang & Liu, 2002; Näther, 2000).

Specifically, by considering the cluster analysis framework, there are different studies regarding partitioning of complex-structured data (i.e. interval valued, symbolic, fuzzy and mixed data).

Different authors suggested conceptual hierarchical and non hierarchical clustering for symbolic data. Michalski & Stepp (1983) developed the algorithm CLUSTER/2, a conjunctive conceptual clustering where descriptive concepts are conjunctive statements involving relations on selected objects features and optimized according to the certain criterion of clustering quality. Chen & Fu (1985) introduced the procedure called HUATUO which produced intermediate conceptual structures for rule-based systems. Fisher (1987), by considering a category utility metric called COBWEB, proposed a top-down incremental conceptual clustering. Ralambondrainy (1995) proposed a conceptual *k*-means clustering method for mixed data (with numerical and symbolic features) based on coding symbolic data numerically and using a mix of Euclidean and Chi-square distances to compute the distance between the hybrid types of data that are represented by considering predicates as groups of attribute-value tuples joined by logical operators.

Diday & Brito (1989) utilized a transfer algorithm for partitioning a set of symbolic data into clusters described by weight distribution vectors. Concerning hierarchical methods, Gowda & Diday (1991, 1992) suggested an agglomerative approach which forms composite symbolic objects utilizing a joint operator whenever mutual pairs of symbolic objects are selected for agglomeration based on minimum dissimilarity (Gowda & Diday, 1991) or maximum similarity (Gowda & Diday, 1992). Gowda & Ravi (1995a,b) introduced, respectively, divisive and agglomerative techniques for symbolic data based on the combined usage of similarity and dissimilarity measures that are defined on the basis of the position, span and content of symbolic objects. Successively, the same authors presented a hierarchical clustering method for symbolic data based on the gravitational approach, which is inspired on the movement of particles in space due to their mutual gravitational attraction (Gowda & Ravi, 1999a) and an ISODATA clustering algorithm for symbolic data using distributed genetic algorithms (Gowda & Ravi, 1999b).

Ichino & Yaguchi (1994) defined generalized Minkowski metrics for mixed feature variables based on the so-called Cartesian space model and presented dendrograms obtained from the application of standard linkage methods for datasets containing numerical and symbolic feature values.

Hathaway et al. (1996) and Pedrycz et al. (1998) introduced models that converted a parametric or nonparametric linguistic variable to generalized coordinates (a vector of numbers) before doing fuzzy *k*-means clustering. Yang & Ko (1996) proposed a class of fuzzy *k*-number clustering procedures for clustering fuzzy data. These procedures have been used to handle certain special types of LR-type fuzzy numbers and also for fuzzy regression analysis (Yang & Ko, 1997). Successively, Yang & Liu (1999) extended the previous work to high-dimensional fuzzy vectors. A fuzzy *k*-means clustering model for symbolic data has been proposed by El-Sonbaty & Ismail (1998a). They computed the distance between the objects from summation of the dissimilarity due to the position, span and content of every attribute. Furthermore, the authors modified the membership and cluster center update equations of the fuzzy *k*-means algorithm to incorporate this dissimilarity measure. Notice that, these authors also suggested an on-line agglomerative hierarchical technique (the single linkage method) for clustering both symbolic and numerical data (El-Sonbaty & Ismail, 1998b). Chavent (2000) proposed a hierarchical divisive clustering method for symbolic data based on a generalized within-cluster inertia criterion. Chavent & Lechevallier (2002) also suggested a *k*-means algorithm for interval valued data. In particular, they proposed a dynamic cluster algorithm for interval valued data where the prototype is defined by the optimization of a criterion based on the Hausdorff distance. Gordon (2000) suggested an iterative relocation algorithm for partitioning symbolic data into classes so as to minimize the sum of the description potentials of the classes. Auephanwiriyakul & Keller (2002) presented in their work a linguistic version of the fuzzy *k*-means method suggested by Bezdek (1981). The suggested algorithm is based on the extension principle and the decomposition theorem. Coppi & D'Urso (2002, 2003) proposed, in a three-way framework, different fuzzy *k*-means clustering models for fuzzy time trajectories that are a particular geometrical representation of the fuzzy data time array. Mali & Mitra (2003) suggested clustering of symbolic data, using different validity indices, for determining the optimal number of meaningful clusters. The novelty of the proposed method lies in transforming the different clustering validity indices, like Normalized Modified Hubert's statistic, Davies-Bouldin index and Dunn's index, from the numerical domain to the symbolic framework. Masson & Denœux (2004) suggested a procedure that generalizes a clustering algorithm based on the belief functions theory introduced by Denœux & Masson (2002) for crisp (non fuzzy) relational data.

Yang et al. (2004) proposed a fuzzy clustering algorithms for mixed features of symbolic and fuzzy data, by modifying Gowda-Diday's dissimilarity measure for symbolic data (Gowda & Diday, 1991, 1992) and also changing the parametric approach for fuzzy data suggested by Hathaway et al. (1996). de Souza & de Carvalho (2004) introduced adaptive and non-adaptive clustering methods for interval valued data based on city-block distances. They suggested two dynamic clustering methods for partitioning a set of symbolic objects where each object is represented by a vector of intervals. The first method utilizes a suitable extension of the city-block distance which compares a pair of vector of intervals. The latter method uses two adaptive versions of this extended city-block distance for interval valued data. In the first version, the adaptive distance has only a single component, whereas it has two components in the second version. In both methods, the prototype of each cluster is also represented by a vector of intervals whose bounds, for each interval, are the median of the set of lower bounds and the median of the set of upper bounds of the intervals of the objects belonging to the cluster (de Souza & de Carvalho, 2004). Guru et al. (2004) suggested a novel similarity measure for estimating the degree of similarity between two patterns, described by interval type data. In particular, this measure computes the degree of similarity between two patterns and approximated the calculated similarity value by a multi-valued type data. Then, based on this similarity, the authors modified the agglomerative method by proposing the concept of mutual similarity value for clustering symbolic pattern. De Carvalho et al. (2004) proposed a dynamic clustering technique for interval valued data based on L_2 distance.

Analogously to non complex datasets, in many real applications, the complex data, i.e. interval valued data, are bound to have noise and outliers. However, in the extensive robust literature (see Section 3), we have not found clustering models for interval valued data that take into account the presence of possible anomalous interval valued data ("noise interval valued data" or "outlier interval valued data"), i.e. interval valued data with anomalous position (location) and/or anomalous shape in the observational space.

In this paper, by considering the fuzzy approach, we suggest a robust *k*-means clustering model for classifying interval valued data. In particular, in Section 2, we introduce briefly the concept of interval valued data and consider a suitable distance measure between interval valued data. In Section 3, we propose a new robust fuzzy clustering model for interval valued data and, successively, in Sections 4 and 5, we show simulative and applicative examples.

2 Distance for interval valued data

In this section we introduce a suitable distance measure between observation units characterized by *p* intervals. Let us indicate the generic interval valued datum pertaining to the *i*-th observation unit with respect to the *j*-th interval valued variable as the couple (c_{ij},r_{ij}), where c_{ij} denotes the center and r_{ij} the radius. The lower and upper bounds of the interval are then obtained as c_{ij}-r_{ij} and c_{ij}+r_{ij}, respectively. If we deal with *p* standard numerical variables, each observation unit is represented as a point in the reference space $\Re^p$. Instead, in case of interval valued data, each observation unit is represented as a hyperrectangle (in $\Re^p$) having 2^p vertices (a rectangle with 2^p =4 vertices if *p*=2).

Several authors propose suitable distances for interval valued data (and, in general, for symbolic data). See, for instance, Gowda & Diday (1991,1992), de Carvalho (1994), Ichino & Yaguchi (1994), Gowda & Ravi (1995a, b), de Carvalho & de Souza (1998) and Chavent & Lechevallier (2002). In this paper, we adopt the distance proposed by D'Urso & Giordani (2004). In order to compare two observation units characterized by *p* interval valued variables, we compare all the vertices of the hyperrectangles pertaining to the observation units involved. We then have:

$$d^2(i',i'')=\sum_{s=1}^{2^p}\left\|(\mathbf{c}_{i'}+\mathbf{r}_{i'}*\mathbf{h}_s)-(\mathbf{c}_{i''}+\mathbf{r}_{i''}*\mathbf{h}_s)\right\|^2 , \qquad (1)$$

where $\mathbf{c}_i$ and $\mathbf{r}_i$ are, respectively, the vectors of the centers and radii of order *p* pertaining to the *i*-th observation unit. The vectors $\mathbf{c}_i$ and $\mathbf{r}_i$ are, respectively, the *i*-th row of **C** (the centers matrix of order *n*×*p*, where *n* denotes the number of observatio units) and **R** (the radii matrix of order *n*×*p*). In (1), the symbol $*$ is the Hadamard product, that is the elementwise product of two matrices (vectors) of the same order. Moreover, the vectors, $\mathbf{h}_s$, $s=1,\dots,2^p$, have elements equal to 1 and –1 and their role is to consider every vertex of the hyperrectangles associated to the observation units. See, for further details, D'Urso & Giordani (2004). It can be shown that the distance in (1) can be simplified as

$$d^2(i',i'')=2^p\left\|\mathbf{c}_{i'}-\mathbf{c}_{i''}\right\|^2+2^p\left\|\mathbf{r}_{i'}-\mathbf{r}_{i''}\right\|^2\approx\left\|\mathbf{c}_{i'}-\mathbf{c}_{i''}\right\|^2+\left\|\mathbf{r}_{i'}-\mathbf{r}_{i''}\right\|^2= \sum_{j=1}^{p}\left[\left(c_{i'j}-c_{i''j}\right)^2+\left(r_{i'j}-r_{i''j}\right)^2\right] \quad (2)$$

It is fruitful to remark that the distance measure in (2) is the same distance used in de Carvalho et al. (2004), which is a special case of the one in de Carvalho & de Souza (1998). The robust fuzzy *k*-means model introduced in the next section aims at clustering observation units described by interval valued variables considering a suitable loss function, which involves the distance in (2).

3 Robust fuzzy clustering for interval valued data set with outliers

In the fuzzy clustering of non interval valued data, the study on the treatment of anomalous data (outlier or noise data) has been widely analyzed (see, for instance, Beni & Liu, 1994; Davè, 1991; Davè & Fu, 1994; Davè & Krishnapuram, 1997; Davè & Sen, 2002; Frigui & Krishnapuram, 1999; Keller, 2000; Kim et al., 1996; Krishnapuram & Keller, 1993, 1996; Ohashi, 1984). In particular, in order to reduce the effect of outliers in the fuzzy clustering, we can consider the following approaches (D'Urso, 2005): *metric approach* (the clustering models belonging to this approach neutralize the disruptive effects of outliers by incorporating, in the objective functions of the clustering models, metrics with robust properties (Kersten, 1999; Hathaway et al., 2000; Leşki, 2003)); *possibilistic approach* (for avoiding the drawback of outliers, the clustering model belonging to this approach considers outliers with small membership degrees to all groups of data (Krishnapuram & Keller, 1993, 1996)); *noise approach* (the models belonging to this approach assign outliers to a special cluster of data (the noise cluster) and reduce the influence of this class on the whole partition (e.g., Davè, 1991; Davè & Krishnapuram, 1997; Davè & Sen, 2002)); *semi-fuzzy approach* (Selim & Ismail (1984) suggest an approach to avoid the inconvenience of outliers in the clustering process, to let a datum belong to a maximum number of clusters, to set the membership degrees to zero if a predefined maximal distance is exceeded, or to define a minimum threshold for the membership degrees); *influence weighting approach* (in the clustering model suggested by Keller (2000) weights for each datum are adapted during the clustering in order to detect whether single data points can be seen as outliers. Also the dynamic fuzzy clustering models with influence weighting system, proposed by D'Urso (2005) for classifying time trajectories in a three-way framework, belong to this approach).

2.1 The model

In this section, following the noise approach, we propose a robust fuzzy *k*-means clustering model for interval valued data. Our model represents an extension of Davè's procedure (1991) -which uses a criterion similar to Ohashi's (1984)- for interval valued data set.

Following the ideas of Ohashi and Davè, by means of the suggested robust fuzzy *k*-means clustering model for interval valued data, we introduce a special cluster, the noise cluster, whose role is to localize the noise and place it in a single auxiliary class. By assigning patterns to the noise class, we declare them to be outliers in the interval valued data set.

Taking into account (2), the suggested robust fuzzy clustering model can be formalized in the following way:

$$\text{minimize: } J_m(\mathbf{U},\overline{\mathbf{C}},\overline{\mathbf{R}};\mathbf{C},\mathbf{R},k)=$$

$$\sum_{i=1}^{n}\sum_{q=1}^{k}u_{iq}^{m}\left[\ \left\|\mathbf{c}_i-\mathbf{c}_q\right\|^2+\left\|\mathbf{r}_i-\mathbf{r}_q\right\|^2\right]+\sum_{i=1}^{n}\delta^2\left(1-\sum_{q=1}^{k}u_{iq}\right)^m \qquad (3)$$

where $\mathbf{U}=\{u_{iq}: i=1,\dots,I; q=1,\dots,k\}$ is the membership degrees matrix whose generic element $u_{iq}\,(\geq 0)$ indicates the membership degree of the *i*-th object to the *q*-th cluster; $\overline{\mathbf{C}}=\{\mathbf{c}_q: q=1,\dots,k\}$ is the center-prototype matrix with generic row $\mathbf{c}_q$ (the *q*-th center prototype); $\overline{\mathbf{R}}=\{\mathbf{r}_q: q=1,\dots,k\}$=radius-prototype matrix with generic row $\mathbf{r}_q$ (the *q*-th radius prototype); $\delta^2>0$ is a suitable scale parameter to be chosen in advance. Such a parameter plays the role to increase (for high values of δ) or to decrease (for low values of δ) the emphasis of the "noise component" of the minimization function in (3).

Notice that, we end up with k+1 clusters, with the extra cluster serving as the noise cluster. The difference in the second term of the objective function $J_m(\mathbf{U},\overline{\mathbf{C}},\overline{\mathbf{R}};\mathbf{C},\mathbf{R},k)$ expresses the degree membership of each pattern to the noise cluster and the sum over the first *k* is lower than or equal to 1. In particular, let the membership degrees u_{i*} of the *i*-th object to the *noise cluster* be defined as $u_{i*}=1-\sum_{q=1}^{k}u_{iq}$. Here, *k* is the number of *good* clusters and u_{iq} denotes the membership degree of the *i*-th observation to the *k*-th fuzzy cluster. Since $u_{i*}=1-\sum_{q=1}^{k}u_{iq}$ is used to define the membership u_{i*} to the *noise cluster*, the usual constraint of the fuzzy *k*-means clustering model ($\sum_{q=1}^{k} u_{iq} = 1$) is not required. Thus, the membership constraint for the *good clusters* is effectively relaxed to $\sum_{q=1}^{k} u_{iq} \leq 1$. This allows noise data to have arbitrarily small membership values in *good clusters*.

The objective function $J_m(\mathbf{U},\overline{\mathbf{C}},\overline{\mathbf{R}};\mathbf{C},\mathbf{R},k)$ can be optimized (minimized) with respect to the center-prototypes, radius-prototypes and membership degrees in a similar manner to the fuzzy *k*-means clustering model proposed by Dunn (1974) and Bezdek (1974, 1981) and then to the noise clustering model suggested by Davè (1991). In particular, the *membership degrees* are:

$$u_{iq} = \frac{1}{\sum_{q'=1}^{k}\left[\frac{\left(\|\mathbf{c}_i-\mathbf{c}_q\|^2+\|\mathbf{r}_i-\mathbf{r}_q\|^2\right)}{\left(\|\mathbf{c}_i-\mathbf{c}_{q'}\|^2+\|\mathbf{r}_i-\mathbf{r}_{q'}\|^2\right)}\right]^{\frac{1}{m-1}}+\left[\frac{\left(\|\mathbf{c}_i-\mathbf{c}_q\|^2+\|\mathbf{r}_i-\mathbf{r}_q\|^2\right)}{\delta^2}\right]^{\frac{1}{m-1}}}. \quad (4)$$

The *center-prototypes* and the *radius-prototypes* are, respectively:

$$\mathbf{c}_q = \frac{\sum_{i=1}^{n} u_{iq}^{m}\,\mathbf{c}_i}{\sum_{i=1}^{n} u_{iq}^{m}}, \quad (5)$$

$$\mathbf{r}_q = \frac{\sum_{i=1}^{n} u_{iq}^{m}\,\mathbf{r}_i}{\sum_{i=1}^{n} u_{iq}^{m}}. \quad (6)$$

For the sake of completeness, notice that, analogously to the non interval case, the selection of δ is a complex issue. After several simulation studies, we observed that, in the interval case, the following value can be chosen :

$$\delta = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{p} c_{ij}}{np}} + \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{p} r_{ij}}{np}}. \quad (7)$$

The suggested model is introduced to make the fuzzy *k*-means clustering model less sensitive to noise and outlier interval valued data by relaxing the constraint on the membership degrees so that the sum of membership degrees of a noise object to all the good classes is not forced to be equal to 1; for this reason it represents a robustified version of the fuzzy *k*-means model for interval valued data and can be easily utilized instead of the fuzzy *k*-means clustering model.

4 Simulative examples

We now propose two applications of our robust fuzzy *k*-means model on simulated data in order to show how the model is able to detect anomalous interval valued data. The two simulated data sets are displayed in Figures 1 and 2. In both cases, n=12 observations are described by p=2 intervals. Two clusters can be easily distinguished. Each cluster is formed by five observations. In particular, the first five observation units pertain to one cluster and the latter five to the other cluster. For their features, two observations (n. 6 and n.7) are anomalous. In the first data set, as one can see from Figure 1, their locations are quite far from both clusters, whereas their shapes are consistent to those of the remaining observation units. In the second data set (given in Figure 2), the shapes corresponding to observation units n.6 and n.7 are bigger than those of the other rectangles, whereas, in this case, the positions of rectangles n.6 and n.7 are not anomalous. In particular, their locations are consistent to observation units n.1-n.5 for observation unit n.6 and to n.8-n.12 for observation unit n.7.

Figure 1: Simulated data set n.1

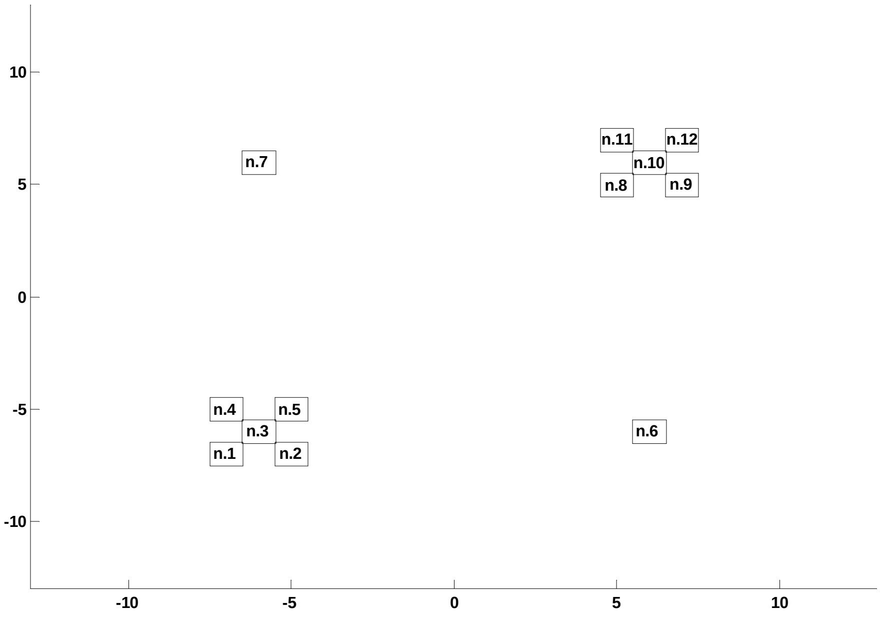
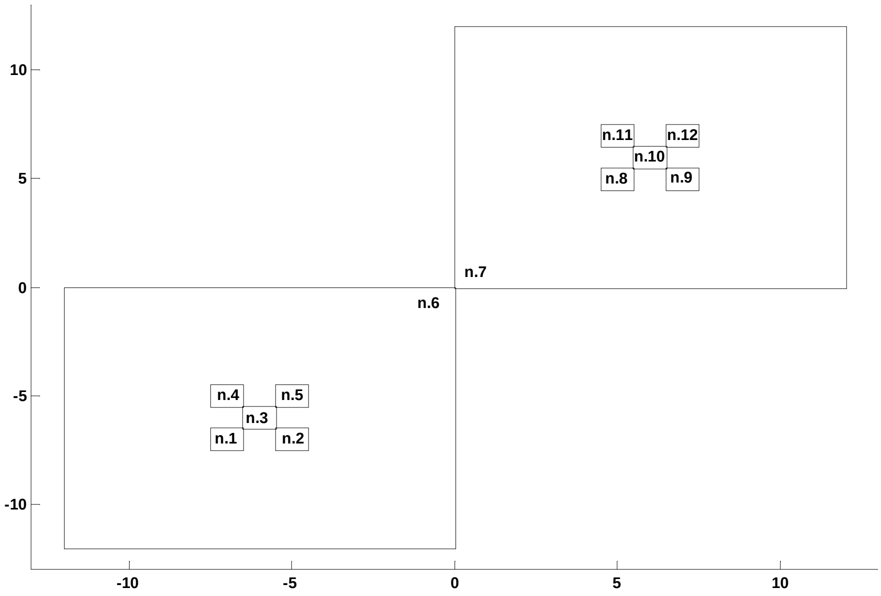


Figure 2: Simulated data set n.2



n.1

By setting *m*=2 and *k*=2, we get the optimal membership degree matrices given in Table 1.

Table 1: Membership degree matrices using the robust fuzzy *k*-means model

Simulated Data Set n.1					Simulated Data Set n.2			
Obs.	Cluster 1	Cluster 2	Noise Cl.		Obs.	Cluster 1	Cluster 2	Noise Cl.
n.1	0.96	0.01	0.03		n.1	0.96	0.01	0.03
n.2	0.95	0.01	0.04		n.2	0.96	0.01	0.03
n.3	1.00	0.00	0.00		n.3	1.00	0.00	0.00
n.4	0.95	0.01	0.04		n.4	0.96	0.01	0.03
n.5	0.94	0.01	0.05		n.5	0.96	0.01	0.03
n.6	0.19	0.19	0.62		n.6	0.53	0.08	0.39
n.7	0.19	0.19	0.62		n.7	0.08	0.53	0.39
n.8	0.01	0.96	0.03		n.8	0.01	0.96	0.03
n.9	0.01	0.95	0.04		n.9	0.01	0.96	0.03
n.10	0.00	1.00	0.00		n.10	0.00	1.00	0.00
n.11	0.01	0.95	0.04		n.11	0.01	0.96	0.03
n.12	0.01	0.94	0.05		n.12	0.01	0.96	0.03

As one may expect, for both data sets, the model correctly detects the membership of the observation units to the clusters. In fact, the first five observations are assigned to the first cluster and the latter five to the second cluster. With regard to the first data set, we can observe that observation units n.6 and n.7 slightly belong to both clusters (membership degree equal to 0.19). Therefore, the robust approach to the clustering problem emphasizes that these two observations ‘pertain’ to what we may call the ‘noise cluster’ (with membership degree equal to 0.62). On the contrary, in the second data set, we can see that observation unit n.6 is partially near to Cluster 1 and observation unit n.7 to Cluster 2. Thus, from Table 1, we can see that observation n.6 is partially assigned to the first cluster (0.53), but the observation also belongs to the ‘noise cluster’, even if to a lesser extent (0.39). The same comment holds with respect to observation unit n.7 and the second cluster.

5 Application

In this section, we provide the results of our robust fuzzy *k*-means model for interval valued data applied to the well-known ‘Fats and Oils’ data set (Ichino & Yabuuchi, 1994). The available data refer to *n*=8 oils described by *p*=4 interval valued variables. For the sake of completeness, we notice that there is also a qualitative variable, which is not considered here. Among the eight fats and oils, six of them are vegetal, whereas two of them (Beef Tallow and Hog Fat) are animal. Among the vegetal oils, it is interesting to notice that two oils are used for paint (Linseed Oil and Perilla Oil), two for foods (Olive Oil and Sesame Oil) and two for cosmetics (Camellia Oil and Cottonseed Oil). It is important to remark that one of the vegetal oils (Linseed Oil) is characterized by anomalous features (especially for the Saponification). We thus expect that our model assigns the Linseed Oil to both the vegetal cluster and to the ‘noise cluster’. Therefore, we decide to apply our model considering *k*=2 clusters. We also set *m*=2. Before performing the model, we preprocess the data by standardizing the centers using the mean and the standard deviation of the original centers and by dividing the radii by the standard deviation of the centers. This way of preprocessing the data helps us to eliminate unwanted differences among the variables, without losing relevant information concerning the width of the intervals.

The membership degree matrix and the centroids (by applying the inverse preprocessing procedure) are given in Tables 2 and 3.

Table 2: Membership degree matrix using the robust fuzzy *k*-means model

Fats and Oils	Cluster 1	Cluster 2	Noise Cluster
Linseed Oil	0.06	0.11	0.83
Perilla Oil	0.08	0.38	0.54
Cottonseed Oil	0.01	0.97	0.02
Sesame Oil	0.01	0.97	0.02
Camellia Oil	0.05	0.75	0.20
Olive Oil	0.04	0.85	0.11
Beef Tallow	0.95	0.01	0.04
Hog Fat	0.96	0.01	0.03

Table 3: Centroids matrix using the robust fuzzy *k*-means model

Centroids	Specific Gravity	Freezing Point	Iodine Value	Saponification
Cluster 1	(0.919,0.002)	(-5.07,2.19)	(102.46,5.45)	(191.49,3.77)
Cluster 2	(0.864,0.004)	(30.17,4.49)	(55.38,8.04)	(195.17,5.31)

The animal fats are assigned to the first cluster with membership degrees equal to 0.95 (Beef Tallow) and 0.96 (Hog Fat). Instead, the vegetal oils pertain to the second cluster. In particular, Cottonseed Oil, Sesame Oil and Olive Oil strongly belong to such a cluster. To a lesser extent, the same comment holds for Camelia Oil (with membership degrees equal to 0.75). Thus, as their membership degrees are rather high, we can conclude that the oils used for foods and cosmetics have almost similar features. Instead, two vegetal oils (Perilla Oil and, especially, Linseed Oil) are assigned to the 'noise cluster'. However, Perilla Oil partially pertains to the second cluster (with membership degree equal to 0.38). By comparing the centroids in Table 3 and the original features of Camellia, we can see similar scores with regard to the Freezing Point and the Saponification Value, whereas the Specific Gravity and the Iodine Value are rather different. The features of Linseed Oil are very different from those, which characterize the two obtained clusters. In fact, all the centers of the four interval valued features are sensibly far from those of the centroids given in Table 3. Moreover, a peculiarity of the Linseed Oil is the anomalous width of the Saponification value: the radius is 39, whereas the radii of the centroids for Saponification are 3.77 (for Cluster 1) and 5.31 (for Cluster 2).

Finally, by observing Table 3, it is interesting to observe that the interval valued centroid pertaining to Cluster 1 is characterized by smaller radii with respect to those of the centroid of the second cluster. The centers of the first centroid are higher than those of the second centroid with regard to the Specific Gravity and the Iodine Value. The opposite comment holds for the Freezing Point and the Saponification.

The obtained clusters are compared to those resulting from the application of the classical fuzzy *k*-means model for interval valued data, as proposed by D'Urso &

Giordani (2005b) [1]. By setting *m*=2 and adopting the same preprocessing procedure, we get the membership degree matrices for *k*=2 and *k*=3 given in Table 4.

Table 4: Membership degree matrix using the classical fuzzy *k*-means model

Fats and Oils	*k*=2 clusters			*k*=3 clusters		
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	Cluster 3
Linseed Oil	0.99	0.01		1.00	0.00	0.00
Perilla Oil	0.27	0.73		0.13	0.13	0.73
Cottonseed Oil	0.02	0.98		0.00	0.01	0.99
Sesame Oil	0.05	0.95		0.00	0.00	1.00
Camellia Oil	0.09	0.91		0.03	0.07	0.90
Olive Oil	0.01	0.99		0.02	0.07	0.91
Beef Tallow	0.14	0.86		0.00	0.99	0.01
Hog Fat	0.13	0.87		0.00	0.99	0.01

In the case *k*=2, we obtain that only the Linseed oil belongs to the first cluster (with membership degree equal to 0.99) and the remaining fats and oils belong to the latter cluster (with membership degrees ranging from 0.73 for the Perilla Oil to 0.99 for the Olive Oil). Thus, in this case, the fat and the oil cannot be distinguished. In case of *k*=3 clusters, we still find that one cluster has only one element (still the Linseed Oil with membership degree equal to 1.00), whereas the remaining two clusters well distinguish the fats and the oils. The Perilla Oil pertains to the cluster of oils with membership degree equal to 0.73. Therefore, the use of the classical fuzzy *k*-means model for interval valued data seems to be inappropriate in the sense that the Perilla Oil and the Linseed Oil are characterized by anomalous features and should not pertain to any of the clusters. In particular, it is strongly inappropriate that Linseed Oil is exactly assigned to one cluster. Moreover, only when *k*=3, the classical fuzzy *k*-means model distinguishes the animal fats and the vegetal oils.

6 Conclusion

In this paper, we have suggested a robust version of the fuzzy *k*-means clustering model for classifying objects with respect to a set of interval valued variables. In

[1] D'Urso & Giordani (2005b) propose a fuzzy *k*-means clustering model for symmetric fuzzy data by introducing a suitable dissimilarity measure for fuzzy data. The dissimilarity measure involved considers the sum of the distance for the centers and the distance for the spreads (the information about the width of a fuzzy datum). These two distance components are differently weighted by means of two weights constructed in such a way that the weight for the centers distance is equal or higher than that for the spreads distance. Such a dissimilarity measure can be also adapted to interval valued data by imposing that the weight for the centers distance is equal to the one for the spreads (radii) distance since the membership function is uniform.

case of interval valued data, the concept of outliers is related to the values of the centers, the radii or both.

We have mathematically formalized the model. Moreover, some illustrative examples, based on synthetic and real data, are considered to show how the suggested model behaves. The clustering model is constructed in such a way that anomalous observation units are assigned to the so-called “noise cluster”.

On the basis of the good results of the application of our model to synthetic and real data, we indicate some possible future perspectives of research in the robust fuzzy clustering framework for interval valued data.

1. Simulation studies for analyzing the computational performances of our clustering model.
2. Cluster-validity criteria for selecting suitably *m* and *k* in the suggested robust fuzzy clustering model.
3. Interval versions of the fuzzy clustering model belonging to other robust approaches (metric, possibilistic, semi-fuzzy and influence weighting approach).

References

Auephanwiriyakul, S., & Keller, J.M., (2002) Analysis and efficient implementation of a linguistic fuzzy *c*-means, IEEE Transactions on Fuzzy Systems, 10 (5), 563-582.

Beni, G., & Liu, X., (1994) A least biased fuzzy clustering method, IEEE Transactions on Pattern Recognition Analysis and Machine Intelligence, 16 (9), 954-960.

Bezdek, J.C., (1974) Numerical taxonomy with fuzzy sets, Journal of Mathematical Biology, 1, 57-71.

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

Billard, L., & Diday, E., (2000) Regression analysis for interval-valued data, in: Data Analysis, Classification and Related Methods (eds. Kiers, H.A.L., Rasson, J.P., Groenen, P.J.F., & Schader, M.), 369-380, Springer, Berlin.

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

Cazes, P., Chouakria, A., Diday, E., & Schektman, Y., (1997) Extension de l’analyse en composantes principales à des données de type intervalle. Revue de Statistique Appliquée, 45, 5-24.

Chavent, M., (2000) Criterion-based divisive clustering for symbolic objects, in Analysis of Symbolic Data (eds. Bock, H.H., & Diday, E.), Springer-Verlag, Heidelberg.

Chavent, M., & Lechevallier, Y., (2002) Dynamical clustering algorithm of interval data: optimization of an adequacy criterion based on Hausdorff distance, in: Classification, Clustering and Data Analysis (eds. Jajuga, K., Sokolowski, A. & Bock, H.H.,), Springer, Heidelberg, 53-59.

Chen, Y., & Fu, K.S, (1985) Conceptual clustering in knowledge organization, IEEE Transactions on Pattern Recognition and Machine Intelligence, 7, 592-598.

Coppi R., & D'Urso, P., (2002) Fuzzy *K*-means clustering models for triangular fuzzy time trajectories, Statistical Methods and Applications, 11 (1), 21-40.

Coppi R., & D'Urso, P., (2003a) Regression analysis with fuzzy informational paradigm: A Least-Squares Approach Using Membership Function Information, International Journal of Pure and Applied Mathematics, 8 (3), 279-306.

Coppi R., & D'Urso, P., (2003b) Three-way fuzzy clustering models for LR fuzzy time trajectories, Computational Statistics & Data Analysis, 43, 149-177.

Coppi, R., D'Urso, P., & Giordani, P., (2005), Component models for fuzzy data, submitted.

D'Urso, P., (2003) Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data, Computational Statistics & Data Analysis, 42 (1-2), 47-72.

D'Urso, P., (2005) Fuzzy clustering for data time arrays with inlier and outlier time trajectories, IEEE Transactions on Fuzzy Systems, in press.

D'Urso, P., & Gastaldi, T., (2000) A least-squares approach to fuzzy linear regression analysis, Computational Statistics & Data Analysis, 34, 427-440.

D'Urso, P., & Gastaldi, T., (2002) An "orderwise" polynomial regression procedure for fuzzy data, Fuzzy Sets and Systems, 130 (1), 1-19.

D'Urso, P., & Giordani, P., (2003) Fitting of fuzzy linear regression models with multivariate response, International Mathematical Journal, 3 (6), 655-664.

D'Urso, P., & Giordani, P., (2004) A least squares approach to principal component analysis for interval valued data, Chemometrics and Intelligent Laboratory Systems, 70, 179-192.

D'Urso, P., & Giordani, P., (2005a) A possibilistic approach to latent component analysis for symmetric fuzzy data, Fuzzy Sets and Systems, 150, 285-305.

D'Urso, P., & Giordani, P., (2005b) A weighted fuzzy *c*-means clustering model for fuzzy data, Computational Statistics & Data Analysis, 2005, in press.

Davè, R., (1991) Characterization and detection of noise in clustering, Pattern Recognition Letters, 12, 657-664.

Davè, R., & Fu, T., (1994) Robust shape detection using fuzzy clustering: practical applications, Fuzzy Sets and Systems, 65, 161-185.

Davè, R., & Krishnapuram, R., (1997) Robust clustering methods: an unified view, IEEE Transactions on Fuzzy Systems, 5 (2), 270-293.

Davè, R., & Sen, S., (2002) Robust fuzzy clustering of relational data, IEEE Transactions on Fuzzy Systems, 10 (6), 713-727.

de Carvalho, F.A.T., (1994) Proximity coefficients between Boolean symbolic objects, in New Approaches in Classification and Data Analysis (eds. Diday, E., Lechevallier, Y., Schader, M., Bertrand, P., & Burtschy, B.,), Springer, Heidelberg, 387-394.

de Carvalho, F.A.T., Brito, F., & Bock, H.H., (2004) Dynamic clustering for interval data based on L_2 distance, Technical Report n. 0437, IAP Statistics Network.

de Carvalho, F.A.T., & de Souza, R.M.C.R., (1998) New metrics for constrained Boolean symbolic objects, in: Proceedings of the Conference on Knowledge Extraction and Symbolic Data Analysis (KESDA ’98). Office for Official Publications of the European Communities, Luxemburg, 175-187.

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

Denœux, T., & Masson, M.H., (2000) Multidimensional scaling of interval-valued dissimilarity data, Pattern Recognition Letters, 21, 83-92.

Denœux, T., & Masson, M.H., (2004) Principal Component Analysis of fuzzy data using autoassociative neural network, IEEE Transactions on Fuzzy Systems, 12 (3), 336-349.

Diday, E., (Ed.) (1988) The Symbolic Approach in Clustering, Classification and related Methods of Data Analysis, Elsevier, Amsterdam.

Diday, E., & Brito, M.P., (1989) Symbolic cluster analysis, in Conceptual and Numerical Analysis of Data (ed. Opitz, O.), 45-84, Springer-Verlag, Heidelberg.

Dunn, J.C., (1974) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Journal of Cybernetics, 3, 32-57.

El-Sonbaty, Y., & Ismail, M.A., (1998a) Fuzzy clustering for symbolic data, IEEE Transactions on Fuzzy Systems, 6 (2), 195-204.

El-Sonbaty, Y., & Ismail, M.A., (1998b) On-line hierarchical clustering, Pattern recognition Letters, 19, 1285-1291.

Fisher, D.H., (1987) Knowledge acquisition via incremental conceptual clustering, Machine Learning, 2, 103-138.

Frigui, H., & Krishnapuram, R., (1999) A robust competitive algorithm with applications in computer vision, IEEE Transactions on Pattern Analysis and Machine Intelligence, 21 (5), 450-465.

Giordani, P., & Kiers, H.A.L., (2004) Principal Component Analysis of symmetric fuzzy data, Computational Statistics & Data Analysis, 45, 519-548.

Gordon, A.D., (2000) An iterative relocation algorithm for classifying symbolic data, in Data Analysis: Scientific Modeling and Practical Application (eds. Gaul, W.; Opitz, O., Schader, M.,) , 17-23, Springer-Verlag, Heidelberg,

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 Transactions on Systems, Man, and Cybernetics, 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 Letters, 16, 647-652.

Gowda, K.C., & Ravi, T.R., (1999a) Clustering of symbolic objects using gravitational approach, IEEE Transactions on Systems, Man, and Cybernetics, 29 (6), 888-894.

Gowda, K.C., & Ravi, T.R., (1999b) An ISODATA clustering procedure for symbolic objects using a distributed genetic algorithm, Pattern Recognition Letters, 20, 659-666.

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 Letters, 25, 1203-1213.

Hathaway, R.J., Bezdek, J.C., & Pedrycz, W., (1996) A parametric model for fusing heterogeneous fuzzy data, IEEE Transactions on Fuzzy Systems, 4 (3), 1277-1282.

Hathaway, R.J., Bezdek, J.C., & Hu, Y., (2000) Generalized fuzzy c-means clustering strategies using L_p norm distances, IEEE Transactions on Fuzzy Systems, 8 (5), 576-582.

Hong, D.H., & Hwang, C., (2005) Interval regression analysis using quadratic loss support vector machine, IEEE Transactions on Fuzzy Systems, 13, 2, 229- 237.

Ichino, M., & Yaguchi, H., (1994) Generalized Minkowsky metrics for mixed feature-type data analysis, IEEE Transaction of Systems, Man and Cybernetics, 24 (4), 698-708.

Kersten, P.R., (1999) Fuzzy order statistics and their application to fuzzy clustering, IEEE Transactions on Fuzzy Systems, 7 (6), 708-712.

Keller, A., (2000) Fuzzy clustering with outliers, in 19th Intern. Conf. of the North American Fuzzy Information Processing Society-NAFIPS “Peach FuzzAtlanta”, 143-147.

Kim, J., Krishnapuram, R., & Davé, R.N., (1996) Application of the least trimmed squares techniques to prototype-based clustering, Pattern Recognition Letters, 17, 633-641.

Körner, R., & Näther, W., (1998) Linear regression with random fuzzy variables: extended classical estimates, best linear estimates, least squares estimates, Inform. Science, 109, 95-118.

Krishnapuram, R., & Keller, J., (1993) A possibilistic approach to clustering, IEEE Transactions on Fuzzy Systems, 1, 98-110.

Krishnapuram, R., & Keller, J., (1996) The possibilistic c-means algorithm: insights and recommendations, IEEE Transactions on Fuzzy Systems, 4, 385-393.

Lauro, C., & Palumbo, F., (2000) Principal component analysis of interval data: a symbolic data analysis approach, Computational Statistics, 15, 73-87.

Lauro, C., & Palumbo, F., (2005) Principal component analysis for non-precise data, in: New Developments in Classification and Data Analysis (eds. Vichi, M., Monari, P., Mignani, S., & Montanari, A.), 173-184, Springer, Berlin.

Lauro, C., Verde, R., & Palumbo, F., (2000) Factorial methods with cohesion constraints on symbolic objects, in: Data Analysis, Classification and Related Methods (eds. Kiers, H.A.L., Rasson, J.P., Groenen, P.J.F., & Schader, M.), 381-386, Springer, Berlin.

Leşki, J.M., (2003) Towards a robust fuzzy clustering, Fuzzy Sets and Systems, 137, 215-233.

Mali, K., & Mitra S., (2003) Clustering and its validation in a symbolic framework, Pattern Recognition Letters, 24, 2367-2376.

Masson, M.-H., & Denœux, T., (2002) Multidimensional scaling of fuzzy dissimilarity data, Fuzzy Sets and Systems, 128 (33), 55-68.

Masson, M.-H., & Denœux, T,. (2004) Clustering interval-valued proximity data using belief functions, Pattern Recognition Letters, 25, 163-171.

Michalski, R., & Stepp, R.E., (1983) Automated construction of classifications: conceptual clustering versus numerica taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence, 5, 396-410.

Näther, W., (2000) On random fuzzy variables of second order and their application to linear statistical inference with fuzzy data, Metrika, 51, 201-221.

Ohashi, Y., (1984) Fuzzy clustering and robust estimation, in 9th Meeting SAS Users Group Int., Hollywood Beach.

Palumbo, F., & Lauro, C., (2003) A PCA for interval-valued data based on midpoints and radii, in: New Developments in Psychometrics (eds. Yanai H., Okada A., Shigemasu K., Kano Y., & Meulman J.,), 641-648, Springer Verlag, Tokyo

Pedrycz, W., Bezdek, J.C., Hathaway, R.J., & Rogers, G.W., (1998) Two nonparametric models for fusing heterogeneous fuzzy data, IEEE Transactions on Fuzzy Systems, 6 (3), 411-425.

Ralambondrainy, H., (1995) A conceptual version of the *K*-means algorithm, Pattern Recognition Letters, 16, 1147-1157.

Selim, S.Z., & Ismail, M.A., (1984) Soft clustering of multidimensional data: a semi-fuzzy approach, Pattern Recognition, 17 (5), 559-568.

Watada, J., & Yabuuchi, Y., (1997) Fuzzy principal component analysis and its application, Biomedical Fuzzy Human Sciences, 3 83-92.

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

Yang, M.S., & Ko, C.H., (1996) On a class of fuzzy *c*-numbers clustering procedures for fuzzy data, Fuzzy Sets and Systems, 84, 49-60.

Yang, M.S., & Liu, H.H., (1999) Fuzzy clustering procedures for conical fuzzy vector data, Fuzzy Sets and Systems, 106, 189-200.

Yang, M.S., & Liu, T.S., (2002) Fuzzy least-squares linear regression analysis for fuzzy input-output data, Fuzzy Sets and Systems, 126, 389-399.