PRZEGLĄD STATYSTYCZNY
R. LXII – ZESZYT 3 – 2015

MARCIN PEŁKA[1], ANDRZEJ DUDEK[2]

THE COMPARISON OF FUZZY CLUSTERING METHODS FOR SYMBOLIC INTERVAL-VALUED DATA

1. INTRODUCTION

In general terms, clustering methods seek to organize certain sets of objects (items) into clusters in the way allowing objects from the same cluster be more similar to each other than to objects from other clusters. Usually such similarity is measured by some distance measure (e.g. Euclidean, Manhattan, etc.). Successful application of these methods has been confirmed in many different areas such as taxonomy, image processing, data mining, etc. In general, clustering techniques can be divided into two groups of methods – hierarchical (agglomerative or divisive) and partitioning (see, e.g. Gordon, 1999; Jain et al., 1999).

In cluster analysis, objects (patterns) are usually described by single-valued variables. This allows representing each object as a vector of qualitative or quantitative measurements, where each column represents a variable.

However, this kind of data representation is too restrictive to cover more complex data. If the uncertainty and/or variability of the data are to be taken into account, variables must assume sets of categories or intervals, including frequencies or weights in some cases.

The discussed data are primarily studied using *Symbolic Data Analysis* (SDA). The main aim of Symbolic Data Analysis is to provide suitable methods for managing aggregated or complex data, described by multi-valued variables, where the cells of a data table contain sets of categories, intervals, or weight (probability) distributions (see e.g. Billard, Diday, 2006; Bock et al., 2000).

Conventional hard clustering methods restrict each object of the data set to exactly one cluster. Fuzzy clustering generates a fuzzy partition using the idea of partial membership, expressed by the degree of membership of each object in a given cluster.

In terms of the real-valued data, Dunn (1973) presented one of the first fuzzy clustering methods applying an adequacy criterion based on the Euclidean distance. Bezdek (1981) generalized this method even further. Diday, Govaert (1977) offered

[1] Wrocław University of Economics, Department of Econometrics and Computer Science, 3 Nowowiejska St., 58-500 Jelenia Góra, Poland, corresponding author – e-mail: marcin.pelka@ue.wroc.pl.

[2] Wrocław University of Economics, Department of Econometrics and Computer Science, 3 Nowowiejska St., 58-500 Jelenia Góra, Poland.

one of the first approaches to use adaptive distances in the partitioning of quantitative data. Gustafson, Kessel (1979) introduced the first adaptive fuzzy clustering, based on a quadric distance defined by fuzzy covariance matrix.

More recently, De Carvalho et al. (2006) introduced fuzzy *c*-means clustering algorithms based on adaptive quadratic distances. These distances can be defined by full as well as diagonal fuzzy covariance matrices (estimated globally), or by diagonal fuzzy covariance matrices (estimated locally for each cluster).

Finite-sample properties of spectral clustering have been theoretically studied by many scientists (see Ng et al., 2002; Shi, Malik, 2002; Meila, Shi, 2001; Chung, 1997; von Luxburg et al., 2005; von Luxburg, 2006; Kannan et al., 2000; Guattery, Miller, 1998; de Sa, 2005). Spectral clustering has the advantage of performing well in the presence of the non-Gaussian clusters. This method is also easy to implement. Furthermore, it is also not a disadvantage for the local minima presence (von Luxburg et al., 2005). Additionally, the convergence of the normalized spectral clustering is less difficult to handle than the unnormalized one (von Luxburg et al., 2005). The results obtained by spectral clustering frequently outperform the traditional approaches (see e.g. von Luxburg, 2006). It is due to the fact that spectral clustering makes no assumptions regarding the form of clusters – it can solve very general clustering problems (von Luxburg, 2006, p. 22).

Spectral clustering, however, has certain disadvantages. It can be quite unstable under different choices of parameters for the neighborhood graphs. Many different kernels can be used, each of them leading to different results (Gaussian kernel is used at most cases) – Karatzoglou (2006) presents the applications of different kernels in spectral clustering.

Another important task is to choose a good σ value for the kernel – in the paper published by Karatzoglou (2006) quite an efficient way for estimating the appropriate σ parameter has been proposed. σ is a scaling parameter which should minimize the sum of inter-cluster distances for a given number of clusters. Usually a heuristic algorithm is used to find the best σ value.

Cominetti et al. (2010) proposed a fuzzy spectral clustering algorithm for complex data, referred to as DifFUZZY – which combines the ideas of fuzzy clustering and spectral clustering. It is applicable to a larger class of clustering problems. DifFUZZY is better than traditional fuzzy clustering algorithms in handling "curved" and elongated data sets or those which contain different dispersion (see Cominetti et al., 2010, p. 1). Moreover, DifFUZZY does not require any prior information on the number of clusters. The algorithm of DifFUZZY may be divided into three main steps: 1) the construction of σ-neighborhood graph using the Euclidean norm, to be followed by applying this graph in determining the number of clusters. 2) computation of auxiliary matrices **W**, **D**, **P**, the definition of which can be intuitively understood in terms of diffusion processes on graphs. Matrix **W** uses the idea of Gaussian kernel. Matrix **D** is defined as a diagonal matrix with diagonal elements equal to $\sum_{j=1}^{N} w_{ij}$ (where w_{ij} are

the elements of matrix **W**). Matrix **P** is calculated from the identity matrix, **W** and **D** matrices, as well as γ_2 which is an internal parameter of DifFUZZY. The default value of this parameter is 0.1. 3) calculation of membership values of even soft points.

Cominetti et al. (2010) showed that the fuzzy spectral algorithm DifFUZZY performs well in a number of data sets (both artificial and real) with sizes ranging from tens to hundreds of data points presenting dimensions as high as hundreds.

Interval-valued variables are needed, e.g. when an object represents a group of individuals and the variables used to describe it need to take the value which expresses the variability inherent in the description of a group. Such data arise in practical situations, e.g. recording monthly interval temperatures at meteorological stations, daily interval stock prices, etc. Another source of interval-valued data is the aggregation of huge databases into a reduced number of groups, the properties of which are described by interval-valued variables.

Construction of clustering methods for interval-valued data must take into account that for this kind of data operations of adding, subtracting, multiplying, squaring, calculation of means or calculation of variance are not defined. In literature of the subject some proposals of construction of similar operations and measures can be found can be found (see Billard, Diday, 2006, p. 69–142) and some author even use achievements of interval algebra (Moore, 1966). The most common approach, although, is to base the clustering algorithm on distance measure dedicated for symbolic Boolean object (see table 1) instead of direct calculation on interval variables.

Symbolic Data Analysis provides a number of fuzzy clustering algorithms for interval-valued data. El-Sonbaty, Ismail (1998) introduced a fuzzy *c*-means algorithm to cluster data on the basis of different types of symbolic variables. Yang et al. (2004) proposed fuzzy clustering algorithms for mixed features of symbolic and fuzzy data. De Carvalho (2007) introduced a fuzzy *c*-means and adaptive fuzzy *c*-means methods for interval-valued data, based on the general form of the Euclidean distance. De Carvalho, Tenório (2010) introduced fuzzy *k*-means clustering algorithms for interval-valued data based on adaptive quadric distances.

However, none of the fuzzy clustering methods for interval-valued data presented so far uses the spectral clustering approach. The spectral clustering algorithm proposed by Ng et al. (2002), based on spectral decomposition of the distance matrix, does not, in fact, represent a new clustering method, but rather a new way of preparing data for the well-known *k*-means method. We propose to adapt this popular way of preparing the inputted data to manage interval-valued symbolic data and then to apply the well-known fuzzy *c*-means clustering algorithm.

The recommended algorithm gives a fuzzy partition and a prototype for each cluster by optimizing an adequacy criterion based on a suitable Euclidean distance.

Simulation studies, with artificial and real data sets, confirm the usefulness of the suggested method when dealing data with different cluster structures, noisy variables and/or outliers.

The presented paper is organized as follows. Section 2 discusses three fuzzy clustering methods for symbolic interval-valued data – the fuzzy *c*-means clustering, the adaptive fuzzy *c*-means clustering, the fuzzy *k*-means clustering and compares them with the proposed spectral fuzzy *c*-means algorithm for clustering symbolic data.

To show the usefulness and stability of the proposed method, section 3 presents the results of evaluation studies with different synthetic interval-valued data sets, as well as the application with real interval-valued data sets. Section 4 offers the concluding remarks.

2. FUZZY CLUSTERING METHODS FOR SYMBOLIC INTERVAL-VALUED DATA

Let $\Omega = \{e_1,\ldots,e_n\}$ be the set of *n* objects (patterns), where each object is indexed by *k* and described by *p* interval-valued variables $\{y_1,\ldots,y_p\}$ where each variable is indexed by *j*. An *interval-valued variable X* (see e.g. Billard, Diday, 2006; Bock et al., 2000) is a correspondence defined from Ω in $\Re$ such that for each $k \in \Omega, X(k) = [a,b] \in \Im$, where $\Im\{[a,b]: a,b \in \Re, a \le b\}$ is the set of closed intervals defined from $\Re$. Each object *k* is represented as a vector of intervals $\mathbf{x}_k = (x_{k1},\ldots,x_{kp})$, where $x_{kj} = [a_{kj}, b_{kj}] \in \Im$. In this paper, an interval data table $\{x_{kj}\}_{n\times p}$ is made up of *n* rows standing for *n* objects and *p* columns representing *p* symbolic variables. Each cell of this data table, called also a symbolic data table or a symbolic data matrix, contains an interval $x_{kj} = [a_{kj}, b_{kj}] \in \Im$. Such a symbolic data matrix serves as input data for the computation of a distance matrix through a distance measure suitable for interval-valued data.

As it has been mentioned in section 1 there are three main fuzzy clustering methods for symbolic interval-valued data (see de Carvalho, 2007; de Carvalho, Tenório, 2010):

1. Fuzzy *c*-means clustering.
2. Adaptive fuzzy *c*-means clustering.
3. Fuzzy *k*-means clustering.

Fuzzy *c*-means clustering for symbolic interval-valued data (IFCM) aims at furnishing the fuzzy partition of a data set and a corresponding set of prototypes, so that criterion W^1 measuring the fitting between clusters and their representatives (prototypes) is locally minimized, which is defined as follows:

$$W^1 = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \sum_{j=1}^{p}\left[(a_{kj}-\alpha_{ij})^2 + (b_{kj}-\beta_{ij})^2\right], \tag{1}$$

where: a_{kj} and b_{kj} represent lower and upper bounds of an interval for an object, whereas α_{ij} and β_{ij} stand lower and upper bounds of an interval for cluster prototype.

Fuzzy *c*-means clustering for the interval-valued data is carried out in the following steps (see de Carvalho, 2007, p. 426):

1. Initialization. Fix number of clusters *c*, $2 \le c < n$, fix fuzzification parameter *m*, $1 < m < \infty$, fix iteration limit *T*, fix $\varepsilon > 0$. Initialize $u_{ik}(k = 1,\ldots,n)$ and $(i = 1,\ldots,c)$ of pattern *k* belonging to cluster P_i so that $u_{ik} \ge 0$ and $\sum_{i=1}^{c} u_{ik} = 1$.

2. $t = 1$.
3. Representation step. Membership degree u_{ik} of pattern k belonging to cluster P_i is fixed. Compute the class prototypes:

$$\alpha_{ij} = \frac{\sum_{k=1}^{n} (u_{ik})^m \cdot a_{kj}}{\sum_{k=1}^{n} (u_{ik})^m}, \quad \beta_{ij} = \frac{\sum_{k=1}^{n} (u_{ik})^m \cdot b_{kj}}{\sum_{k=1}^{n} (u_{ik})^m}. \tag{2}$$

4. Allocation step. The prototypes $\mathbf{g}_i$ of class P_i are fixed. Update the fuzzy membership degree as follows:

$$u_{ik} = \left[\sum_{h=1}^{c} \left(\frac{\sum_{j=1}^{p} \left[(a_{kj} - \alpha_{ij})^2 + (b_{kj} - \beta_{ij})^2 \right]}{\sum_{j=1}^{p} \left[(a_{kj} - \alpha_{hj})^2 + (b_{kj} - \beta_{hj})^2 \right]} \right)^{\frac{1}{m-1}} \right]^{-1}. \tag{3}$$

5. Stopping criterion. If $\left| W_{t+1}^1 - W_t^1 \right| \leq \varepsilon$ or $t > T$ then stop, else $t = t + 1$ and go to step 3 (representation step).

Adaptive fuzzy *c*-means clustering for the interval-valued data (IFCMADS) has the same purpose as fuzzy *c*-means clustering for symbolic interval-valued data, but criterion W^2 is defined as follows:

$$W^2 = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} \lambda_{ij} \left[(a_{kj} - \alpha_{ij})^2 + (b_{kj} - \beta_{ij})^2 \right], \tag{4}$$

where: a_{kj} and b_{kj} represent lower and upper bounds of an interval for an object, whereas α_{ij} and β_{ij} stand for lower and upper bounds of an interval for cluster prototype. λ_{ij} are the weights defined by equation 5.

The adaptive fuzzy *c*-means clustering algorithm is carried out based on the following steps (see de Carvalho, 2007, p. 427):

1. Initialization. Fix c, $2 \leq c < n$, fix m, $1 < m < \infty$, fix an iteration limit T, fix $\varepsilon > 0$. Initialize u_{ik} in the same manner as in the IFCM clustering algorithm.
2. $t = 1$.
3. Representation step:
   Stage 1: Membership degree u_{ik} is fixed. Compute cluster prototypes in the same way as in IFCM clustering algorithm.
   Stage 2: Membership degree u_{ik} is fixed and class prototypes are fixed. Compute the vector of weights $\boldsymbol{\lambda}_i$ as follows:

$$\lambda_{ij} = \frac{\left\{ \prod_{h=1}^{p} \left[\sum_{k=1}^{n} (u_{ik})^m \left((a_{kh} - \alpha_{ih})^2 + (b_{kh} - \beta_{ih})^2 \right) \right] \right\}^{\frac{1}{p}}}{\sum_{k=1}^{n} (u_{ik})^m \left((a_{kj} - \alpha_{ij})^2 + (b_{kj} - \beta_{ij})^2 \right)}. \tag{5}$$

4. Allocation step. Class prototypes and the vector of weights λ_i are fixed. Update the fuzzy membership degree u_{ik} of pattern k belonging to cluster P_i as follows:

$$u_{ik} = \left[\sum_{h=1}^{c} \left(\frac{\sum_{j=1}^{p} \lambda_{ij} \left[\left(a_{kj} - \alpha_{ij} \right)^2 + \left(b_{kj} - \beta_{ij} \right)^2 \right]}{\sum_{j=1}^{p} \lambda_{ij} \left[\left(a_{hj} - \alpha_{hj} \right)^2 + \left(b_{kj} - \beta_{hj} \right)^2 \right]} \right)^{\frac{1}{m-1}} \right]^{-1}. \tag{6}$$

5. Stopping criterion. If $\left| W_{t+1}^2 - W_t^2 \right| \leq \varepsilon$ or $t > T$ then stop, else $t = t + 1$ and go to step 3 (representation step).

Fuzzy *k*-means clustering algorithms for the interval-valued data are based on adaptive quadric distances. Fuzzy *k*-means clustering algorithms optimize an adequacy criterion J measuring the fit between clusters and their prototypes, which is defined as (de Carvalho, Tenório, 2010, p. 2980):

$$J = \sum_{k=1}^{K} \sum_{i=1}^{n} \left(u_{ik} \right)^m d^2_{\mathbf{M}_k (\mathbf{x}_i, \mathbf{y}_k)}, \tag{7}$$

where $d^2_{\mathbf{M}_k (\mathbf{x}_i, \mathbf{y}_k)} = \left(\mathbf{x}_{iL} - \mathbf{y}_{kL} \right)^T \mathbf{M}_k \left(\mathbf{x}_{iL} - \mathbf{y}_{kL} \right) + \left(\mathbf{x}_{iU} - \mathbf{y}_{kU} \right)^T \mathbf{M}_k \left(\mathbf{x}_{iU} - \mathbf{y}_{kU} \right)$ is a suitable adaptive quadric distance between vectors of intervals parameterized by a positive definite symmetric matrix $\mathbf{M}_k$, $\mathbf{x}_{iL}, \mathbf{y}_{kL} \left(\mathbf{x}_{iU}, \mathbf{y}_{kU} \right)$ represent lower (L) and upper (U) bounds of intervals.

There are three types of adaptive quadric distances possible to apply:

– Single adaptive quadric distances defined by a full pooled fuzzy covariance matrix $\mathbf{M}_k = \mathbf{M}$.
– Single adaptive quadric distances defined by a diagonal pooled fuzzy covariance matrix $\mathbf{M}_k = \mathbf{M} = Diag\left(\lambda^1, \ldots, \lambda^p \right)$.
– Cluster adaptive quadric distances defined by a full fuzzy covariance matrix.

The fuzzy *k*-means clustering algorithm, regardless of the distance type, is executed in four main steps (de Carvalho, Tenório, 2010, p. 2981–2982; 2983–2984):

1. Initialization. Fix c (number of clusters), $2 \leq c < n$, fix m, $1 < m < \infty$, fix an iteration limit T, fix $\varepsilon > 0$. Initialize u_{ik} in the same manner as in IFCM or IFCMADS algorithms.
2. $t = t + 1$.
3. Definition of the best prototypes, u_{ik} and the corresponding vector of matrices $\mathbf{\Theta} = \{\mathbf{M}_1, \ldots, \mathbf{M}_k\}$ are fixed. Compute the vector of prototypes in the same way as in IFCM algorithm.
4. Definition of the best distances. Both u_{ik} and the corresponding vector of prototypes are fixed. Compute the vector of positive definite symmetric matrices $\mathbf{\Theta}$.
5. Definition of the best partition. The vector of prototypes and the corresponding matrices of weights $\mathbf{\Theta}$ are fixed. Determine the fuzzy partition represented by u_{ik}:

$$u_{ik} = \left[\sum_{h=1}^{K} \left(\frac{d^2_{\mathbf{M}_k}(\mathbf{x}_i, \mathbf{y}_k)}{d^2_{\mathbf{M}_k}(\mathbf{x}_i, \mathbf{y}_k)} \right)^{\frac{1}{m-1}} \right]^{-1}. \tag{8}$$

6. Stopping criterion. If $\left| J_{t+1} - J_t \right| \leq \varepsilon$ or $t > T$ then stop, else $t = t + 1$ and go to step 3.

Spectral fuzzy c-means algorithm covers two steps:

a) spectral decomposition algorithm adapted to deal with an interval-valued data, where $\mathbf{E}'$ or $\mathbf{E}$ matrix is obtained,

b) fuzzy c-means clustering built upon the $\mathbf{E}'$ or $\mathbf{E}$ matrix.

Spectral decomposition algorithm, adapted to deal with an interval-valued data set, takes the following steps (Ng et al., 2002):

1. Let $\mathbf{X}$ be a symbolic data table with n rows and p columns and let c be the number of clusters.
2. Let $\mathbf{S} = [s_{kl}]$ be a similarity matrix between the objects belonging to $\mathbf{X}$. The similarity matrix can be computed using the below equation:

$$s_{kl} = \frac{d_{kl}}{e^{\sigma^2}}, \tag{9}$$

where d_{kl} is a suitable dissimilarity measure computed on the pair of vectors of intervals and σ is a scaling parameter that should minimize the sum of inter-cluster distances for a given number of clusters. Usually a heuristic algorithm is used to find the best σ value.

3. From the similarity matrix $\mathbf{S} = [s_{kl}]$ compute the matrix of weights $\mathbf{W} = [w_{kl}]$ as follows:

$$w_{kl} = \begin{cases} \sum_{l=1}^{n} d^2_{kl} & k \neq l, \\ 0 & k = l. \end{cases} \tag{10}$$

4. Then compute the Laplacian $\mathbf{L}$ matrix according to:

$$\mathbf{L} = W^{-\frac{1}{2}} \times \mathbf{S} \times \mathbf{W}^{-\frac{1}{2}}. \tag{11}$$

In the graph theory, $\mathbf{L}$ is treated as the algebraical representation of the graph created from the objects of $\mathbf{X}$.

5. Extract the first c eigenvectors of the Laplacian matrix to create the matrix $\mathbf{E} = [e_{kl}]$. Each eigenvector of $\mathbf{L}$ is a column of $\mathbf{E}$ (thus, matrix $\mathbf{E}$ is $n \times c$ dimensional). Alternatively, instead of matrix $\mathbf{E}$, the normalized matrix $\mathbf{E}' = [e'_{kl}]$ can be considered, which is computed as follows:

$$e'_{kl} = \frac{E_{kl}}{\sqrt{\sum_{i=1}^{n} E^2_{il}}}. \tag{12}$$

6. Finally, a standard clustering algorithm is applied on matrix $\mathbf{E}'$ or $\mathbf{E}$, if the normalization step is omitted to obtain a suitable clustering structure. In the presented paper, the well-known fuzzy *c*-means represents the considered standard clustering algorithm.

The variants of spectral clustering may differ depending on the applied kernel estimator type. Usually the Gaussian estimator, based on the squared Euclidean distance, is used (for classical data with a ratio and interval variables). For the purposes of interval-valued symbolic data it is suggested to apply the Gaussian estimator based on squared dissimilarity functions that that are suitable for interval-valued data (see Billard, Diday, 2006; Bock et al., 2000; Zelnik-Manor, Perona, 2004). Table 1 presents all suitable distance measures for symbolic interval-valued data available in R software. All of them (except `C_1`, which is the distance measure for hierarchical or logical dependent symbolic variables) will be used in evaluation studies (see section 3).

Parameter σ represents the key element in spectral clustering. There are many heuristic approaches allowing the selection of the best σ value (see e.g. Fisher, Poland, 2004; Poland, Zeugmann, 2006; Zelnik-Manor, Perona, 2004). Parameter σ can be selected using some descriptive statistics computed from the distance matrix. A better way for selecting it was proposed by Karatzoglou (2006) following which a σ that minimizes the total within the sum of squares of distances, computed between the objects for given u clusters, is searched for.

Fuzzy *c*-means clustering algorithm (FCM), proposed by Dunn (1973) and improved by Bezdek (1981), is a very well-known algorithm commonly applied to pattern recognition tasks. FCM clustering algorithm is based on the minimization of the following objective function:

$$J_m = \sum_{k=1}^{n}\sum_{i=1}^{C} u_{ik}^{m}\left\|\mathbf{x}_k - \mathbf{c}_i\right\|^2. \tag{13}$$

FCM algorithm has the following steps:

1. Initialize the membership degrees u_{ik} and form the fuzzy partition matrix $\mathbf{U}^{(0)} = [u_{ik}]$.
2. At the r-th step – compute the center vectors $\mathbf{c}_i^{(r)}$, with $\mathbf{U}^{(r-1)}$ kept fixed, as follows:

$$\mathbf{c}_i^{(r)} = \frac{\sum_{k=1}^{N}\left(u_{ik}^{(r-1)}\right)^m \cdot \mathbf{e}_k}{\sum_{k=1}^{N}\left(u_{ik}^{(r-1)}\right)^m}. \tag{14}$$

3. Compute $\mathbf{U}^{(r)}$, with $\mathbf{c}_i^{(r)}$ kept fixed, as follows:

$$u_{ik}^{(r-1)} = \frac{1}{\sum_{j=1}^{C}\left(\frac{\left\|\mathbf{e}_k - \mathbf{c}_k\right\|}{\left\|\mathbf{e}_i - \mathbf{c}_j\right\|}\right)^{\frac{2}{m-1}}}, \tag{15}$$

where: $\mathbf{e}_k$ and $\mathbf{e}_i$ are k-th and i-th elements of $\mathbf{E}'$ matrix.

4. If $\left\|\mathbf{U}^{(r)} - \mathbf{U}^{(r-1)}\right\| < \varepsilon$ then stop, else return to step 2.

Table 1.

Distance measures for Boolean symbolic objects available in R software

DistType & distance name	Elements of distance measure	Distance measure
`U_2` Ichino-Yaguchi	$\phi(v_{ij}, v_{kj}) = \lvert v_{ij} \oplus v_{kj} \rvert - \lvert v_{ij} \otimes v_{kj} \rvert + \gamma\left(2 \cdot \lvert v_{ij} \oplus v_{kj} \rvert - \lvert v_{ij} \rvert - \lvert v_{kj} \rvert\right)$	$\sqrt[q]{\sum_{j=1}^{m} \phi(v_{ij}, v_{kj})^q}$
`U_3` normalized Ichino-Yaguchi	$\psi(v_{ij}, v_{kj}) = \dfrac{\phi(v_{ij}, v_{kj})}{\lvert V_j \rvert}$	$\sqrt[q]{\sum_{j=1}^{m} \psi(v_{ij}, v_{kj})^q}$
`U_4` weighted and normalized Ichino-Yaguchi	$\phi(v_{ij}, v_{kj})$ same as in `U_2`	$\sqrt[q]{\sum_{j=1}^{m} w_j \psi(v_{ij}, v_{kj})^q}$
`SO_2` de Carvalho	$\psi(v_{ij}, v_{kj}) = \dfrac{\phi(v_{ij}, v_{kj})}{\mu(v_{ij} \oplus v_{kj})}$ $\phi(v_{ij}, v_{kj})$ same as in `U_2`	$\sqrt[q]{\sum_{j=1}^{m} \dfrac{1}{m} \left[\psi(v_{ij}, v_{kj})\right]^q}$
`SO_1` de Carvalho	$\alpha = \mu(v_{ij} \cap v_{kj})$, $\beta = \mu\left[v_{ij} \cap c(v_{kj})\right]$, $\chi = \mu\left[c(v_{ij}) \cap v_{kj}\right]$, $\delta = \mu\left[c(v_{ij}) \cap c(v_{kj})\right]$ $d_1 = \dfrac{\alpha}{\alpha + \beta + \chi}$,	$\sqrt[q]{\sum_{j=1}^{m} \left[w_j d_f(v_{ij}, v_{kj})\right]^q}$
`C_1` de Carvalho for hierarchical or logical dependent variables	$d_2 = \dfrac{2\alpha}{2\alpha + \beta + \chi}$, $d_3 = \dfrac{\alpha}{\alpha + 2(\beta + \chi)}$, $d_4 = \dfrac{1}{2}\left[\dfrac{\alpha}{\alpha + \beta} + \dfrac{\alpha}{\alpha + \chi}\right]$, $d_5 = \dfrac{\alpha}{\sqrt{(\alpha + \beta) + (\alpha + \chi)}}$ $d_f(v_{ij}, v_{kj}) = 1 - D_f$ $f = 1, \ldots, 5$	$\sqrt[q]{\dfrac{\sum_{j=1}^{m} \left[w_j d_f(v_{ij}, v_{kj})\right]^q}{\sum_{j=1}^{m} \delta(V_j)}}$

Table 1. (cont.)

DistType & distance name	Elements of distance measure	Distance measure
SO_3 de Carvalho	–	$[\pi(A_i \oplus A_k) - \pi(A_i \otimes A_k) + \gamma(2\pi(A_i \oplus A_k) - \pi(A_i) - \pi(A_k))]$
SO_4 normalized de Carvalho	–	$[\pi(A_i \oplus A_k) - \pi(A_i \otimes A_k) + \gamma(2\pi(A_i \oplus A_k) - \pi(A_i) - \pi(A_k))] / \pi(A^E)$
SO_5 normalized de Carvalho	–	$[\pi(A_i \oplus A_k) - \pi(A_i \otimes A_k) + \gamma(2\pi(A_i \oplus A_k) - \pi(A_i) - \pi(A_k))] / \pi(A_i \oplus A_k)$
H Hausdorff	–	$\left[\sum_{j=1}^{m} \left(\max\left\{\left\|\overline{v}_{ij} - \overline{v}_{kj}\right\|, \left\|\underline{v}_{ij} - \underline{v}_{kj}\right\|\right\}\right)^2\right]^{\frac{1}{2}}$
L_1	a) interval-valued variables $L_1(v_{ij}, v_{kj}) = \left\|\overline{v}_{ij} - \overline{v}_{kj}\right\| + \left\|\underline{v}_{ij} - \underline{v}_{kj}\right\|$ $L_2(v_{ij}, v_{kj}) = \left\|\overline{v}_{ij} - \overline{v}_{kj}\right\|^2 + \left\|\underline{v}_{ij} - \underline{v}_{kj}\right\|^2$ b) multinominal variables: $L_1(v_{ij}, v_{kj}) = \sum_{y_f} \left\|q_i v_j(y_f) - q_k v_j(y_f)\right\|$ $L_2(v_{ij}, v_{kj}) = \sum_{y_f} \left\|q_i v_j(y_f) - q_k v_j(y_f)\right\|^2$	$\sqrt[q]{\sum_{j=1}^{m} (L_q(v_{ij}, v_{kj}))^q}$ $q = 1$ for L_1 $q = 2$ for L_2
L_2		

Where $v_{ij}v_{kj}$ – realizations of symbolic variables (interval-valued or multinominal), $A_i = (v_{i1},v_{i2},\ldots,v_{im})$ and $A_k = (v_{k1},v_{k2},\ldots,v_{km})$ – i-th and k-th symbolic object described by m symbolic variables, γ – parameter from the range of [0;1], usually $\gamma = \frac{1}{2}$, $q = \{1,2\ldots\}$ (usually $q = 2$), $|\ |$ – for interval-valued data it is the length of the interval, for other variables it is the number of elements, w_j – weight for j-th variable, μ – interval length for interval-valued variables, $c(v_{ij})$ – complement of the symbolic variable V_j, α, β, χ, δ – agreement and disagreement measures for symbolic variables, $\pi(A_i)$ – description potential of i-th symbolic object, A^E – maximum symbolic object according to the descriptive potential, $\delta(V_i)$ – indicator function. It equals 1 when the variable is defined according to logical or hierarchical dependencies with other variables. It equals 0 in other cases. For L_1 and L_2 distance measures in the case of multinominal variables: q.

Source: Gatnar, Walesiak (2011, p. 20–23).

3. EVALUATION EXPERIMENTS

For the purposes of simulation study, four different data sets were prepared with the application of `cluster.Gen` and `genRandomClust` functions of `clusterSim` (Walesiak, Dudek, 2014) and `clusterGeneration` (Qiu, Joe, 2006) packages of `R` software. Models contain the known structure of clusters. Simulation models, generated following the application of `cluster.Gen` function differ in the number of true variables, the density of cluster shapes, the number of true clusters, the number of noisy variables.

In case of symbolic data the data sets can have different shapes – more or less spherical, rounded and non-classical – like smiley, worms, or cuboids that are well-known from `mlbench` package of R software. The general shape of symbolic interval-valued data mimics the desired shape (e.g. spheres, worms, cuboids, smiley, etc.).

In order to obtain the symbolic interval-valued variables, the data were generated twice for each model into sets A and B, while the minimal (maximal) value of $\left\{x_{ij}^{A}, x_{ij}^{B}\right\}$ is treated as the beginning (the end) of an interval. The noisy variables are simulated independently, based on the uniformly distributed random variables. The variances of noisy variables, in the generated data sets, are required to be similar to non-noisy variables (see Milligan, Cooper, 1988; Qiu, Joe, 2006, p. 322).

The models generated by `genRandomClust` function represent data sets with the specified degree of separation (see Qiu, Joe, 2006; Qiu, Joe, 2006a). They differ in the number of true variables, the density and the shapes of clusters, the number of true clusters, the number of noisy variables. In order to build interval data – the obtained data is treated as the center of rectangle. The width and the height of the rectangle are drawn randomly within the of [1, 8].

Real data sets were also used to check the proposed method – well-known Ichiono's oils (Ichino, 1998), cars (de Carvalho et al., 2006) and the European Union countries (Dudek, 2013) data sets.

3.1. ARTIFICIAL DATA SETS

Four different artificial models are used:

1. **Model I**. The model is generated with the application of `clusterSim` package. It contains five clusters in 2 dimensions which are not well separated. The observations are independently drawn from a bivariate normal distribution with means (5,5), (–3,3), (3,–3), (0,0), (–5,–5) and the identity covariance matrix $\sum\left(\sigma_{jj}=1, \sigma_{jl}=-0.9\right)$.
2. **Model II**. The model is generated using `clusterSim` package. It contains five clusters in 3 dimensions which are not well separated. The observations are independently drawn from a multivariate normal distribution with means (5,5,5), (–3,3,–3), (3,–3,3), (0,0,0), (–5,–5,–5) and a covariance matrix $\sum$, where $\sigma_{jj}=1\left(1 \le j \le 3\right)$ and $\sigma_{jl}=0.9\left(1 \le j \ne l \le 3\right)$.

3. **Model III**. The model is generated with the application of `clusterGeneration` package. It also contains five clusters in five dimensions which are not well separated. The desired value of the separation index between (see Qiu, Joe, 2006, Qiu, Joe, 2006a) a cluster and its nearest neighboring cluster was equal to 0.03 and method to generate the covariance matrices for clusters was set to "`onion`".
4. **Model IV**. Model generated with application of `clusterGeneration` package. The model contains six clusters in four dimensions representing overlapping clusters. The desired value of separation index (see Qiu, Joe 2006; Qiu, Joe, 2006a) between a cluster and its nearest neighboring cluster was equal to 0.013 and the method responsible for generating the covariance matrices for clusters was set to "`c-vine`".

For each model 20 simulation runs, with different distance types, were performed. The mean (MR) and the standard deviation (SD) of the fuzzy variant of Rand Index, proposed by Hüllermeier, Rifqi (see Hüllermeier, Rifqi, 2009, p. 1296–1297), were calculated for these trials. The fuzzy variant of Rand index is calculated as follows (see Hüllermeier, Rifqi, 2009, p. 1296–1297):

$$Rand = 1 - dist(P,Q), \tag{16}$$

where: $dist(P,Q)$ – the distance on two fuzzy partitions P and Q defined on the normalized sum of degrees of discordance, calculated as follows:

$$d(P,Q) = \frac{\sum_{(x,x')\in C} \left| E_P(x,x') - E_Q(x,x') \right|}{n(n-1)/2}, \tag{17}$$

$$E_P = 1 - \left\| P(x) - P(x') \right\|, \tag{18}$$

where: $\|\bullet\|$ is the proper distance measure on [0,1], $P_i(x)$ is the membership degree of x in i-th cluster.

Obviously, there are many other validity indices to be used in terms of the results of fuzzy clustering method – see for example Wang, Zhang (2007).

The results of these simulations for different distance measures, applicable for symbolic interval-valued data, are presented in the table 2. Then 20 simulations were also performed for each model with outliers and noisy variables. The mean (MR) and the standard deviation (SD) of the fuzzy version of Rand Index were also computed for these trials – the results are presented in table 3.

Table 2.

The results of simulations for models without noisy variables or outliers for spectral fuzzy *c*-means, fuzzy *c*-means, adaptive fuzzy *c*-means and fuzzy *k*-means clustering

	Model I	Model II	Model III	Model IV
Distance measure	Spectral fuzzy *c*-means			
U_2	MR = 1 SD = 5.078e-13	MR = 1 SD = 4.023e-12	MR = 0.9999 SD = 2.150e-07	MR = 1 SD = 4.054e-06
U_3 & U_4 with equal weights	MR = 0.9997 SD = 6.092e-10	MR = 0.9999 SD = 7.324e-11	MR = 0.9999 SD = 9.221e-11	MR = 0.9999 SD = 4.994e-08
SO_2	MR = 0.9999 SD = 5.003e-11	MR = 0.9999 SD = 5.000e-10	MR = 0.9999 SD = 3.098e-10	MR = 0.9997 SD = 2.208e-11
SO_1 with d_1	MR = 1 SD = 6.005e-10	MR = 0.9996 SD = 4.004e-09	MR = 0.9994 SD = 8.003e-07	MR = 0.9998 SD = 7.011e-07
SO_1 with d_2	MR = 0.9994 SD = 4.213e-10	MR = 0.9921 SD = 6.437e-09	MR = 0.9970 SD = 8.478e-07	MR = 0.9977 SD = 1.987e-07
SO_1 with d_3	MR = 0.9999 SD = 6.827e-09	MR = 0.9967 SD = 8.748e-08	MR = 0.9998 SD = 4.874e-07	MR = 0.990 SD = 8.885e-06
SO_1 with d_4	MR = 0.9998 SD = 4.667e-06	MR = 0.9987 SD = 4.009e-04	MR = 0.9990 SD = 5.330e-05	MR = 0.9986 SD = 5.087e-05
SO_1 with d_5	MR = 0.9900 SD = 4.123e-05	MR = 0.9954 SD = 1.100e-05	MR = 0.9950 SD = 2.083e-05	MR = 0.9903 SD = 7.078e-04
SO_3	MR = 0.9999 SD = 1.913e-06	MR = 0.9999 SD = 4.878375e-08	MR = 0.9999 SD = 7.115587e-08	MR = 1 SD = 5.976251e-08
SO_4	MR = 0.9999 SD = 2.398e-06	MR = 0.9999 SD = 3.866e-07	MR = 0.9999 SD = 6.989e-06	MR = 0.9999 SD = 4.948e-05
SO_5	MR = 0.9999 SD = 5.108e-06	MR = 0.9999 SD = 3.006e-06	MR = 0.9999 SD = 6.072e-05	MR = 0.9999 SD = 6.089e-05
H	MR = 1 SD = 3.113e-08	MR = 1 SD = 3.410e-08	MR = 1 SD = 4.058e-08	MR = 0.9999 SD = 7.485e-08
L_1 and L_2	NA NA	NA NA	NA NA	NA NA
Fuzzy *c*-means				
	MR =1 SD = 8.769e-13	MR = 1 SD = 3.004e-11	MR = 0.9999 SD = 0.0470	MR = 0.8916 SD = 0.0428
Adaptive fuzzy *c*-means				
	MR = 1 SD = 1.152e-11	MR = 1 SD = 5.326e-11	MR = 0.9999 SD = 0.0344	MR = 0.8747 SD = 0.0426
Fuzzy *k*-means				
	MR = 0.9999 SD = 7.944e-07	MR = 0.9999 SD = 4.595e-07	MR = 0.9999 SD = 0.0004	MR = 0.9994 SD = 0.0005

Where MR – mean fuzzy Rand index, SD – standard deviation of fuzzy Rand index, NA – value could not be calculated.

Source: authors' compilation.

Table 3.

The results of simulations for models with noisy variables and/or outliers for spectral fuzzy *c*-means, fuzzy *c*-means, adaptive fuzzy *c*-means, fuzzy *k*-means clustering

	Model I +1 noisy variable & 25% outliers	Model II +45% outliers	Model III +2 noisy variables	Model IV +1 noisy variable & 25% outliers
Distance measure	Spectral fuzzy *c*-means			
U_2	MR = 1 SD = 8.637e-12	MR = 1 SD = 6.463e-13	MR = 0.9953 SD = 0.0003	MR = 0.9837 SD = 0.0016
U_3 & U_4 with equal weights	MR = 0.9986 SD = 4.493e-10	MR = 0.9989 SD = 5.424e-09	MR = 0.9945 SD = 1.091e-10	MR = 0.9960 SD = 6.900e-07
SO_2	MR = 0.9999 SD = 3.472-07	MR = 0.9999 SD = 4.982e-11	MR = 0.9999 SD = 1.387e-10	MR = 0.9999 SD = 3.478e-06
SO_1 with d_1	MR = 1 SD = 3.389e-11	MR = 0.9987 SD = 1.137e-10	MR = 0.9967 SD = 3.873e-10	MR = 0.9940 SD = 5.839e-07
SO_1 with d_2	MR = 0.9973 SD = 4.325e-10	MR = 0.9910 SD = 6.434e-09	MR = 0.9949 SD = 8.532e-05	MR = 0.9956 SD = 1.133e-07
SO_1 with d_3	MR = 0.9999 SD = 3.764e-08	MR = 0.9967 SD = 7.837e-07	MR = 0.9998 SD = 1.387e-06	MR = 0.9900 SD = 3.424e-05
SO_1 with d_4	MR = 0.9987 SD = 2.228e-05	MR = 0.9968 SD = 5.576e-05	MR = 0.9976 SD = 3.347e-06	MR = 0.9950 SD = 4.437e-05
SO_1 with d_5	MR = 0.9967 SD = 3.887e-06	MR = 0.9917 SD = 3.001e-06	MR = 0.9933 SD = 1.378e-05	MR = 0.9900 SD = 2.873e-05
SO_3	MR = 0.9999 SD = 8.551e-07	MR = 0.9999 SD = 4.197e-07	MR = 0.9999 SD = 3.313e-07	MR = 0.9999 SD = 3.422491e-07
SO_4	MR = 0.9997 SD = 4.766e-05	MR = 0.9998 SD = 6.616e-06	MR = 0.9998 SD = 5.428e-05	SD = 0.9999 SD = 7.774e-05
SO_5	MR = 0.9987 SD = 2.135e-05	MR = 0.9988 SD = 8.663e-05	MR = 0.9998 SD = 4.849e-05	MR = 0.9998 SD = 8.117e-05
H	MR = 0.9999 SD = 8.530e-08	MR = 0.9999 SD = 0.021e-07	MR = 0.999 SD = 1.076e-07	MR = 0.9999 SD = 8.592e-07
L_1 and L_2	NA NA	NA NA	NA NA	NA NA
Fuzzy *c*-means				
	MR = 0.7551 SD = 0.2048	MR = 0.7916 SD = 0.0637	MR = 0.8521 SD = 0.0042	MR = 0.6732 SD = 0.3093
Adaptive fuzzy *c*-means				
	MR = 1 SD = 1.136e-03	MR = 0.8934 SD = 4.273e-04	MR = 0.9107 SD = 2.235e-06	MR = 0.9999 SD = 0.0034
Fuzzy *k*-means				
	MR = 0.9983 SD = 0.0004	MR = 0.9978 SD = 7.500e-05	MR = 0.9994 SD = 2.342e-04	MR = 0.9972 SD = 5.345e-04

Where: all elements are the same as in table 2.

Source: authors' compilation.

3.2. REAL DATA SETS

A car symbolic interval data set consists of 33 objects (car models) described by 8 interval-valued variables, 2 categorical multi-nominal variables and one nominal variable (de Carvalho et al., 2006). In this application, only 8 interval-valued variables – *Price*, *Engine Capacity*, *Top Speed*, *Acceleration*, *Step*, *Length*, *Width* and *Height* – were considered for clustering purposes. This data set was clustered 20 times into 4 clusters using all of the applicable distances. The best results were obtained for unnormalized Ichino and Yaguchi, Hausdorff and normalized Ichino and Yaguchi distance measures. The mean Rand index for fuzzy data for the normalized Ichino and Yaguchi distance equals 0.9999983, the standard deviation is 1.384948e-06. For the Hausdorff distance the mean Rand index equals 1, whereas its standard deviation is 0, the same result is achieved in case of Ichino and Yaguchi distance.

The Ichino's oils data set consists of 8 oils and fats described by 8 interval--valued variables (Ichino, 1988) – *Specific gravity*, *Freezing point*, *Iodine value* and *Saponification value*. This data set was 20 times clustered into 2 clusters by applying all distances. The best results were obtained for Hausdorff, Ichino and Yaguchi and De Carvalho distances. In case of Hausdorff and the De Carvalho distances the mean Rand index equals 1 and its standard deviation is 0 or nearly 0. In case of Ichnino and Yaguchi distance the mean Rand index equals 0.9898985 and its standard deviation is 0.03109184.

The European Union (EU) – data set consists of 27 European Union countries described by 8 interval-valued variables representing innovation indicators within the EU countries (Dudek 2013) – *R & D expenditures*, *Enterprises with innovation activity*, *Expenditures on education*, *Internet access*, *Patents per million of citizens*, *e-Administration accessibility indicator*, *Broadband Internet*, *High-technology trade (exports)*. In case of De Carvalho, Ichino and Yaguchi and the normalized Ichino and Yaguchi distance measures the mean Rand index equals 1 (or nearly 1) and its standard deviation is 3.88068e-07.

4. FINAL REMARKS

The main contribution of this paper is the introduction of spectral fuzzy *c*-means algorithm (SCFM) for the symbolic interval-valued data. Due to the fact that the discussed algorithm is based on spectral decomposition of the distance matrix, it can be easily applied to any other symbolic data types and selecting the suitable distance remains the only requirement. SCFM starts from the symbolic data matrix followed by a distance matrix calculation. Spectral decomposition of the distance table is performed, and then the well-known fuzzy *c*-means algorithm is applied.

The spectral fuzzy *c*-means clustering requires the selection of σ parameter for the kernel, which can turn out difficult. However, the solution proposed by Karatzoglou can be used along with the selection of distance measure for symbolic data. Experiments

show that Hausdorff (H) distance reaches the best results (in terms of Rand index mean) when dealing with data sets without noisy variables and outliers.

When clustering symbolic data with (or without) noisy variables and/or outliers SCFM with the application of Hausdorff (`H`) and De Carvalho (`SO_3`) distances generally reach better results than SCFM with the application of normalized Ichino and Yaguchi (`U_3`, `U_4`) distances. Unnormalized Ichino and Yaguchi (`U_2`) distance sometimes reaches similar results as Hausdorff and De Carvalho distance measures. Slightly worse results are reached for the data sets with noisy variables than for the data sets with outliers. The above result was expected due to the fact that this method is based on distance measurements. It can be omitted by using some sort of variable selection algorithm, e.g. HINoV for symbolic data, which is available in `clusterSim` package (see Walesiak, Dudek, 2014; Walesiak, Dudek, 2008), or Ichino and Yaguchi feature selection for symbolic data (see Dudek et al., 2014) available in `symbolicDA` package of R software.

The spectral fuzzy c-means clustering, due to spectral decomposition of data matrix, can deal quite easily with data sets that have some "non-classical" shapes, known from `mlbench` package of R software, like cubes, worms, etc. Furthermore, if the number of clusters is exceeding the actual number of clusters in the data set, the membership degree for those exceeding clusters is getting lower and lower by each iteration.

Experiments with artificial data sets, with different cluster structures or the set degree of cluster separation, without noisy variables and/or outliers, show that this method reaches quite stable results – in terms of fuzzy Rand index. The same results appear while dealing with real data sets and artificial data sets with noisy variables and/or outliers.

One problem (limitation) appears while attempting to apply `L_1`-type distance measure to symbolic interval-data with low degrees of cluster separation. SCFM could not calculate eigenvectors due to the fact that, in this case, objects are too close to each other. The only solution in such a situation, i.e. while trying to apply this method and this distance measure, is to add some slight noise. It will not change the cluster structure, however, calculating eigenvalues will be possible.

When compared to other fuzzy clustering methods for symbolic data the proposed methods usually offer quite good results (in terms of Rand index mean and its standard deviation).

REFERENCES

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

Billard L., Diday E., (2006), *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, Wiley, Chichester.

Bock H.-H., Diday E. (eds.), (2000), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin-Heidelberg.

Chung F., (1997), *Spectral graph theory*, Washington, Conference Board of the Mathematical Sciences.

Cominetti O., Matzavinos A., Samarasinghe S., Kulasiri D., Maini P. K., Erban R., (2010), DifFUZZY: A Fuzzy Spectral Clustering Algorithm For Complex Data Sets, *International Journal of Computational Intelligence in Bioinformatics and Systems Biology*, 1 (4), 402–417.

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 Letters*, 27 (3), 167–179.

De Carvalho F. A. T., Tenório C. P., Cavalcanti Junior N. L., (2006), Partitional Fuzzy Clustering Methods Based On Adaptive Quadratic Distances, *Fuzzy Sets and Systems*, 157, 2833–2857.

De Carvalho F. A. T., (2007), Fuzzy C-means Clustering Methods For Symbolic Interval Data, *Pattern Recognition Letters*, 28 (4), 423–437.

De Carvalho F. A. T., Tenório C. P., (2010), Fuzzy K-means Clustering Algorithms For Interval-valued Data Based On Adaptive Quadric Distances, *Fuzzy Sets and Systems*, 161 (23), 2978–2999.

de Sa V. R., (2005), *Spectral Clustering With Two Views*, ICML Workshop on Learning with Multiple Views.

Diday E., Govaert G., (1977), Classification Automatique Avec Distances Adaptatives, *R.A.I.R.O. Informatique Computer Science*, 11 (4), 329–349.

Dunn J. C., (1973), A Fuzzy Relative ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, *Journal of Cybernetics*, 3, 32–57.

Dudek A., (2013), *Metody analizy danych symbolicznych w badaniach ekonomicznych*, Wrocław University of Economics Publishing House, Wrocław.

Dudek A., Pełka M., Wilk J., (2014), The `symbolicDA` package, `http://www.R-project.org`.

El-Sonbaty Y., Ismail M.A., (1998), Fuzzy Clustering For Symbolic Data, *IEEE Transactions on Fuzzy Systems*, 6, 195–204.

Fischer I., Poland J., (2004), *New methods for spectral clustering*, Technical Report No. IDSIA-12-04, Dalle Molle Institute for Artificial Intelligence, Manno-Lugano, Switzerland.

Gatnar E., Walesiak M., (eds.), (2011), *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, C.H. Beck, Warszawa.

Gordon A. D., (1999), *Classification*, Chapman and Hall/CRC, Boca Raton.

Guattery S., Miller G.L., (1998), On the Quality of Spectral Separators, *SIAM Journal on Matrix Analysis and Applications*, 19 (3), 701–719.

Gustafson D. E., Kessel W. C., (1979), *Fuzzy Clustering with Fuzzy Covariance Matrix*, Proceedings of IEEE Conference on Decision and Control, San Diego, CA, 761–766.

Hüllermeier E., Rifqi M., (2009), *A Fuzzy Variant of the Rand Index for Comparing Clustering Structures*, Proceedings of IFSA/EUSFLAT Conference '2009, 1294–1298.

Ichino M., (1988), *General Metrics for Mixed Features – The Cartesian Space Theory for Pattern Recognition*, Proceedings of the 1988 IEEE International Conference on Systems, Man and Cybernetics, 1, 494–497, International Academic Publishers Beijing.

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

Kannan R., Vempala S., Vetta A., (2000), *On Clusterings – Good, Bad and Spectral*, Technical Report, Computer Science Department, Yale University.

Karatzoglou A., (2006), *Kernel Methods. Software, Algorithms and Applications*, Doctoral thesis, Vienna University of Technology.

Malerba D., Esposito F., Gioviale V., Tamma V., (2001), *Comparing Dissimilarity Measures for Symbolic Data Analysis*, Pre-Proceedings of ETK-NTTS 2001, Hersonissos, 473-48.

Meila M., Shi J., (2001), *A Random Walks View of Spectral Segmentation*, 8-th International Workshop on Artificial Intelligence and Statistics (AISTATS).

Milligan G. W., Cooper M. C., (1988), A Study of Standardization of Variables in Cluster Analysis, *Journal of Classification*, 5 (2), 181–204.

Moore R..E., (1966), *Interval Analysis*, Prentice-Hall, Englewood Cliffs, NJ.

Ng A., Jordan M., Weiss Y., (2002), On Spectral Clustering: Analysis and Algorithm, in: Dietterich T., Becker S., Ghahramani Z., (eds.), *Advances in Neural Information Processing Systems*, 14, MIT Press, 849–856.

Nieddu L., Rizzi A., (2005), Metrics in Symbolic Data Analysis, in: Vichi M., Monari P., Signani S., Montanari A., (eds.), *New Development in Classification and Data Analysis*, Springer-Verlag, Berlin-Heidelberg, 71–78.

Poland J., Zeugmann T., (2006), Clustering the Google Distance with Eigenvectors and Semidefinite Programming, in: Jantke K. P., Kreuzberger G., (eds.), *Diskussionsbeiträge*, *Institut für Medien und Kommunikationswisschaft*, Technische Universität Ilmenau, 21, 61–69, July 2006.

Shi J., Malik J., (2000), Normalized Cuts and Image Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (8), 888–905.

Qiu W., Joe H., (2006), Generation of Random Clusters With Specified Degree of Separation, *Journal of Classification*, 23 (2), 315–334.

Qiu W., Joe H., (2006a), Separation Index and Partial Membership for Clustering, *Computational Statistics and Data Analysis*, 50, 585–603.

Qiu, W., Joe, H. (2010), The `clusterGeneration` package, `http://www.R-project.org`.

von Luxburg U., Bousquet O., Belkin M., (2005), *Limits of Spectral Clustering*, in: Saul L., Weiss Y., Bottou L., (eds.), *Advances in Neural Information Processing Systems (NIPS)* 17, Cambridge, MA: MIT Press, 857–864.

von Luxburg U., (2006), *A Tutorial on Spectral Clustering*, Max Planck Institute for Biological Cybernetics, Technical Report TR-149.

Walesiak M., Dudek A., (2008), Identification of Noisy Variables for Nonmetric and Symbolic Data in Cluster Analysis, in: Preisach C., Burkhardt H., Schmidt-Thieme L., Decker R., (eds)., *Data Analysis, Machine Learning and Applications*, Springer-Verlag, Berlin-Heidelberg, 85–92.

Walesiak M., Dudek A., (2014), The `clusterSim` package, `http://www.R-project.org`.

Wang W., Zhang Y., (2007), On Fuzzy Validity Indices, *Fuzzy Sets and Systems*, 158, 2095–2117.

Zelnik-Manor L., Perona P., (2004), *Self-tuning Spectral Clustering*, Proceedings of the 18th Annual Conference on Neural Information Processing Systems (NIPS'04), `http://books.nips.cc/nips17.html`.

Yang M.-S., Hwang P.-Y., Chen D.-H., (2004*),* Fuzzy Clustering Algorithms for Mixed Feature Types, *Fuzzy Sets Systems*, 141, 301–317.

Yaguchi H., Ichino M., (1994), Feature Selection for Symbolic Data Classification, in: Diday E. Lechevallier Y., Schader M., Bertrand P., Burtschy B., (eds.), *New Approaches in Classification and Data Analysis*, Springer-Verlag, Berlin-Heidelberg, 387–394.

PORÓWNANIE METOD KLASYFIKACJI ROZMYTEJ DLA DANYCH SYMBOLICZNYCH INTERWAŁOWYCH

Streszczenie

Dane symboliczne interwałowe mogą znaleźć zastosowanie w wielu sytuacjach – np. w przypadku notowań giełdowych, zmianach kursów walut, itp. Celem artykułu jest porównanie trzech metod klasyfikacji rozmytej dla danych symbolicznych interwałowych – tj. rozmytej klasyfikacji *c*-średnich, adaptacyjnej rozmytej klasyfikacji *c*-średnich oraz rozmytej klasyfikacji *k*-średnich z rozmytą klasyfikacją spektralną. Rozmyta klasyfikacja spektralna stanowi połączenie podejścia spektralnego oraz klasyfikacji rozmytej *c*-średnich, dzięki czemu możliwe jest otrzymanie lepszych rezultatów (w sensie indeksu Randa dla klasyfikacji rozmytych). Przeprowadzone badania symulacyjne wskazują, że rozmyta klasyfikacja spektralna dla danych symbolicznych pozwala na uzyskanie lepszych wyników niż inne rozmyte metody

klasyfikacji dla tego typu danych jeżeli weźmiemy pod uwagę zbiory danych o różnej strukturze klas, która dodatkowo jest zniekształcana przez obserwacje odstające lub zmienne zakłócające.

Słowa kluczowe: klasyfikacja spektralna, klasyfikacja rozmyta, dane symboliczne interwałowe, analiza danych symbolicznych

THE COMPARISON OF FUZZY CLUSTERING METHODS FOR SYMBOLIC INTERVAL-VALUED DATA

Abstract

Interval-valued data can find their practical applications in such situations as recording monthly interval temperatures at meteorological stations, daily interval stock prices, etc. The primary objective of the presented paper is to compare three different methods of fuzzy clustering for interval-valued symbolic data, i.e.: fuzzy *c*-means clustering, adaptive fuzzy *c*-means clustering and fuzzy *k*-means clustering with fuzzy spectral clustering. Fuzzy spectral clustering combines both spectral and fuzzy approaches in order to obtain better results (in terms of Rand index for fuzzy clustering). The conducted simulation studies with artificial and real data sets confirm both higher usefulness and more stable results of fuzzy spectral clustering method, as compared to other existing fuzzy clustering methods for symbolic interval-valued data, when dealing with data featuring different cluster structures, noisy variables and/or outliers.

Keywords: spectral clustering, fuzzy clustering, fuzzy partition, interval-valued data, symbolic data analysis