Clustering of modal-valued symbolic data

Nataša Kejžar1 · Simona Korenjak-Černe2 · Vladimir Batagelj3,4,5

Received: 12 August 2014 / Revised: 20 August 2020 / Accepted: 12 October 2020 / Published online: 24 October 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract
Symbolic data analysis is based on special descriptions of data known as symbolic objects (SOs). Such descriptions preserve more detailed information about units and their clusters than the usual representations with mean values. A special type of SO is a representation with frequency or probability distributions (modal values). This representation enables us to simultaneously consider variables of all measurement types during the clustering process. In this paper, we present the theoretical basis for compatible leaders and agglomerative clustering methods with alternative dissimilarities for modal-valued SOs. The leaders method efficiently solves clustering problems with large numbers of units, while the agglomerative method can be applied either alone to a small data set, or to leaders, obtained from the compatible leaders clustering method. We focus on (a) the inclusion of weights that enables clustering representatives to retain the same structure as if clustering only first order units and (b) the selection of relative dissimilarities that produce more interpretable, i.e., meaningful optimal clustering representatives. The usefulness of the proposed methods with adaptations was assessed and substantiated by carefully constructed simulation settings and demonstrated on three different real-world data sets gaining in interpretability from the use of weights (population pyramids and ESS data) or relative dissimilarity (US patents data).

Keywords Symbolic objects · Leaders method · Hierarchical clustering · Ward’s method · Clustering demographic structures · United States Patents data set · European social survey data set

Mathematics Subject Classification 62H30 · 91C20 · 62-07 · 68T10

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11634-020-00425-4) contains supplementary material, which is available to authorized users.

Extended author information available on the last page of the article
1 Introduction

In traditional data analysis a unit is usually described with a list of numerical, ordinal or nominal values of selected variables. In symbolic data analysis (SDA) a unit of a data set can be represented, for each variable, with a more detailed description than only a single value. Such structured descriptions are usually called symbolic objects (SOs) (Bock and Diday 2000; Billard and Diday 2006). A special type of SO consists of descriptions with frequency or probability distributions. In this way, we can simultaneously consider both single-value variables and variables with richer descriptions. Computerization of data gathering worldwide has resulted in enormous data sets. The predefined aggregation (preclustering) of raw data is becoming a common method to preserve as much information as possible from this type of data.

For example, if a large store chain that records each purchase of its customers may want information about patterns of customer purchases, a likely technique would be to aggregate customer purchases inside a selected time window. One customer variable can be the yearly shopping pattern for a selected item. Such a variable could be described with a single number (average yearly purchase) or with a symbolic description (e.g., purchases on that item aggregated by month). The second description is richer and enables better analyses.

We adapted two classical clustering methods to retain and use more information about each unit during the clustering process:

– leaders method, which is a generalization of the k-means method (Hartigan 1975; Anderberg 1973; Diday 1979)
– Ward’s hierarchical clustering method (Ward 1963).

Both methods are compatible, i.e., they are both based on the same criterion function. Therefore, they solve essentially the same clustering optimization problem. They can be used in combination: the leaders method can reduce the size of the set of units to a manageable number of leaders that can be further clustered using the compatible hierarchical clustering method. This process helps revealing the relations among the clusters/leaders and deciding, using the dendrogram, upon an appropriate number of final clusters.

Since clustering objects into similar groups plays an important role in exploratory data analysis, many clustering approaches have been developed in SDA to compare SOs based on different dissimilarities. They can be found in books and papers from the field: Bock and Diday (2000), Billard and Diday (2003, 2006, 2019), Diday and Noirhomme-Fraiture (2008) and Noirhomme-Fraiture and Brito (2011). Although most attention has been focused on clustering interval data (de Carvalho et al. 2006) some methods, similar to our approach, have been developed for modal-valued data: Gowda and Diday (1991), Ichino and Yaguchi (1994); Verde et al. (2000), Korenjak-Černe and Batagelj (1998), Korenjak-Černe and Batagelj (2002), Irpino and Verde (2006) and Verde and Irpino (2010).

The paper Hardy and Lallemand (2004) is focused on the detection of the correct number of clusters. Their study is based on an adapted clustering method for symbolic data, SCLUST (Symbolic dynamic CLUSTering method, which is based on the definition of dissimilarity, where weights are not included), and on classical
hierarchical methods (single linkage, complete linkage, centroid, and Ward), where only dissimilarities are adapted for the symbolic multi-valued data ($L_1$, $L_2$ and Carvalho distance). No weights are used. The paper of Irpino et al. (2014) describes a dynamic clustering approach for histogram data based on the Wasserstein distance. This distance also enables automatic computation of relevance weights for variables. The approach is appealing but cannot be used when clustering general (not necessarily numerical) modal-valued data. In their paper de Carvalho and Sousa (2010) present an approach with dynamic clustering that could be (with a pre-processing step) used to cluster any type of symbolic data. The adaptive squared Euclidean distance is used for the dissimilarities. The number of clusters has to be determined in advance when using only dynamic clustering. This can be avoided by combining dynamic clustering with compatible hierarchical methods. In their paper, Kim and Billard (2011) propose the Ichino-Yaguchi dissimilarity measure extended to histogram data, and in Kim and Billard (2012), two measures (Ichino-Yaguchi and Gowda-Diday) are extended to general modal-valued data. The authors use these measures with a divisive clustering algorithm and propose two cluster validity indexes to determine the optimal number of final clusters. In the paper Kim and Billard (2013), even more general dissimilarity measures are proposed to use with mixed histogram, multi-valued and interval data.

Our work is directly motivated by clustering several real-world data sets, e.g., demographic structures or households, which by themselves represent so-called second order units, or patent citations, where we want to obtain informative optimal citation patterns of patents in the same cluster. The main research question of our work is how to define dissimilarities to obtain interpretable (meaningful) optimal cluster leaders. In this paper, we provide the theoretical basis for the generalization of compatible leaders and agglomerative hierarchical clustering methods for modal-valued data that would yield such informative/meaningful cluster leaders. The novelty in our paper is the adaptation of an agglomerative hierarchical and leaders criterion function that can be used with alternative dissimilarity measures and that allows for the use of weights for each SO (or even its variable’s component) to consider the size (counts/frequencies of values, totals, areas and similar—depending on the data) of each SO. Each of the alternative dissimilarities can be used in the leaders and agglomerative hierarchical clustering method, thus enabling the user to chain both methods. When analyzing large data sets, we can use the leaders method to shrink the data set into a more manageable number of clusters (each represented by its leader), which can be further clustered via hierarchical method. Thus, the number of final clusters is finally determined from the dendrogram.

The following problems (which motivated our work) can occur when clustering units described with frequency distributions:

**Problem 1:** The values in the descriptions of different variables can be based on different numbers of original units.

Descriptions with distributions allow for the combination of different—related data sets with distinct micro units. The question arises whether each micro unit should contribute the same amount of information to the result or should less importance be given to a single micro unit if agglomerated units consist of a larger number of micro units. One possible approach to address this problem is presented in the application in
Korenjak-Černe et al. (2011), where two related data sets (teachers and their students) are combined in an ego-centered network, which is presented with a symbolic data description.

**Problem 2:** The representative of a cluster does not preserve the information about the variables’ distributions and the number of first order units when clustering aggregated data units (the representatives are not real structures of the first order units, i.e., they are not meaningful).

Cluster representatives are descriptions of final clusters, optimized by the selected criterion. The optimal representatives from the classical average or barycentric solution are only technical constructs and do not have straightforward interpretations. In some applications, one would like the representatives to present real structures of first order units (i.e., frequency or probability distributions). For example, this problem appears when the clustering units are the age-sex structures of the world’s countries (e.g., Irpino and Verde 2006; Košmelj and Billard 2011). In Korenjak-Černe et al. (2015), the authors used a weighted agglomerative clustering approach to cluster population pyramids of the world’s countries, where cluster representatives are real age-sex structures.

**Problem 3:** The squared Euclidean distance does not give very informative clusterings when the empirical probability distributions (of modal-valued variables) have high peaks.

In the clustering of citation patterns, Kejžar et al. (2011) showed that since the Euclidean distance favors modal values, the optimal cluster leaders show patterns with only one peak that do not properly represent the citation patterns of the patents in the cluster. The use of a relative error measure (relative dissimilarity) is suggested to avoid this problem. In some cases when relative error measures are used also count zeros have to be tackled.

In this paper, we show that these three problems can be solved using the generalized leaders and agglomerative methods with appropriate selections of dissimilarity and weights. These methods produce more meaningful cluster representatives.

In the following section, we introduce the notation and present the development of the adapted methods. The third section illustrates the usefulness of the proposed methods by simulations, and in the fourth section, we apply the methods on three real-world data sets. Section five concludes the paper. In the Supplementary material, we provide the proofs that all the alternative dissimilarities can also be used with the proposed approach. We also provide more elaborate results of the proposed methods applied to the simulated and the real-world data sets.

## 2 Clustering

The set of units \( U \) consists of SOs. An SO \( X \) is described with a list of descriptions of variables \( V_i, i = 1, \ldots, m \)

\[ x = [f_{x_1}, f_{x_2}, \ldots, f_{x_m}] \]
In our general model, each variable is described with a list of values $f_{x_i}$. Note that $f_{x_i}$ is an abbreviation for $(f_x)_i$. The number of variables is denoted by $m$, and

$$f_{x_i} = [f_{x_{i1}}, f_{x_{i2}}, \ldots, f_{x_{ik_i}}],$$

where $k_i$ is the length (the number of terms) of the list for $f_{x_i}$ of variable $V_i$, and $f_{x_{ij}}$ is the absolute frequency/quantity of category $j$ for variable $V_i$ for SO $X$. Variable $V_i$ has $k_i$ categories.

Let $n_{x_i}$ be the sum of frequencies/quantities of a variable $V_i$

$$n_{x_i} = \sum_{j=1}^{k_i} f_{x_{ij}}.$$

Then, we obtain the corresponding empirical probability distribution

$$p_{x_i} = \frac{1}{n_{x_i}} f_{x_i}.$$

In general, we can transform a variable measured in any traditional measurement scale (nominal, ordinal, numerical) into a nominal variable by partitioning its range (set of possible values) into a smaller number of subsets—categories. In this case, some information is lost (i.e., ordinality). However, more variability is preserved, and such a description enables us to simultaneously consider variables of all types.

The leaders method and the agglomerative hierarchical clustering method are two approaches to solving the clustering optimization problem. We use the criterion function of the following form

$$P(C) = \sum_{C \in \mathbf{C}} p(C).$$

The total error $P(C)$ of the partition $C$ is the sum of cluster errors $p(C)$ of its clusters $C \in \mathbf{C}$.

There are many ways to measure the cluster error $p(C)$. In this paper, we shall assume a model in which the error of a cluster is the sum of deviations of its units from the cluster’s representative $T$. For a given representative $T$ and a cluster $C$ we define the cluster error with respect to $T$:

$$p(C, T) = \sum_{X \in C} d(X, T),$$

where $d$ is a selected dissimilarity measure. The best representative $T_C$ is called a leader

$$T_C = \arg \min_T p(C, T).$$
Then, we define

\[ p(C) = p(C, T_C) = \min_T \sum_{X \in C} d(X, T) . \]  

(1)

We assume that the leader \( T \) has the same description structure as the SOs (i.e., it is represented with a list of nonnegative vectors \( t_i \) of size \( k_i \) for each variable \( V_i \)). We do not require that they are distributions; therefore, the representation space is \( \mathcal{T} = (\mathbb{R}^+_0)^{k_1} \times (\mathbb{R}^+_0)^{k_2} \times \cdots \times (\mathbb{R}^+_0)^{k_m} \).

We introduce a dissimilarity measure between SOs and \( T \) with

\[ d(X, T) = \sum_i a_i d_i(x_i, t_i), \quad a_i \geq 0, \quad \sum_i a_i = 1, \]

where \( a_i \) are the weights for variables that could be introduced by the user (i.e., to be able to specify the importance of variables based on external knowledge). We do not consider them further. The dissimilarities \( d_i \) have similar form

\[ d_i(x_i, t_i) = \sum_{j=1}^{k_i} w_{x_i,j} \delta(p_{x_i,j}, t_{ij}), \quad w_{x_i,j} \geq 0 \]

where \( w_{x_i,j} \) are the weights for each variable’s component. Using an alternative basic dissimilarity \( \delta \), we can address problem 3. Six examples of basic dissimilarities \( \delta \) are presented in Table 1. The table lists the basic dissimilarities between the unit’s component \( x \) and the leader’s component \( t \), as proposed in Kejžar et al. (2011) for the classical data representation. In this paper, we extend them to modal-valued SOs. All dissimilarities but \( \delta_1 \) extend the squared Euclidean distance in such a way to assure that the dissimilarities between \( x \) and \( t \) do not increase as a quadratic function. Measures \( \delta_k, k \in \{3, \ldots, 6\} \) currently appear to be less attractive for applications, especially from the implementation point of view (they have some limitations regarding the empirical probability distributions of units, they are restricted to positive \( p_{x_i,j} \), see Sect. 2.1.2). We nevertheless performed simulation experiments with them and analyzed the data sets of world population pyramids and US patents and obtained sensible results (see Supplement, Sects. S3.2, S4.2). When using \( \delta_1 \), the dissimilarity \( d(X, T) \) defined above is a generalization of the squared Euclidean distance.

For the same unit \( X \), the weight \( w_{x_i,j} \) can be different for each variable \( V_i \) and for each of its components. With weights \( w \), we can take into account a different number of original units for each variable in the clustering process (solving problem 1 and/or problem 2), and they also allow regulation of the importance of each variable’s category. For example, the population pyramid of a country \( X \) can be represented with two symbolic variables (one for each gender), where people of each gender are represented with the distribution over age groups. Here, \( w_{x_1,j} = n_{x_1} \) (for all \( j \in \{1, \ldots, k_1\} \)) is the number of all men (the weight is the same for every age group), and \( n_{x_2} \) is the number of all women in country \( X \).

To include and preserve the information about the variables’ distributions and their size throughout the clustering process (problem 2), the following has to hold for a general clustering method when merging two disjoint clusters \( C_u \) and \( C_v \) into a new
cluster $C_z$ (clusters $C_u$ and $C_v$ may consist of only a single unit):

$$p_{zi} = \frac{1}{W_{zi}} f_{zi}$$

where $p_{zi}$ denotes the relative frequency distribution of the variable $V_i$ of the joint cluster $C_z = C_u \cup C_v$, $f_{zi}$ is the frequency distribution of variable $V_i$ in the joint cluster and $w_{zi} = w_{ui} + w_{vi}$ is the weight of that variable in the joint cluster.

Although we use the notation $f$ for modal data, which usually implies absolute frequencies, other interpretations of $f$ and $w$ are possible. For example, $w$ can represent the money spent, and $f$ can represent the distribution of the money spent on a selected item within a given time period. We could even consider information from other sources to determine weights; however, we do not explore this further.

### 2.1 Leaders method

The leaders method, also called the dynamic clustering method (Diday 1979), is a generalization of a popular nonhierarchical $k$-means clustering method (Anderberg 1973; Hartigan 1975). The idea is to obtain the optimal clustering into a pre-specified number of clusters with an iterative procedure. For a current clustering, the leaders are determined as the best representatives of its clusters, and the new clustering is determined by assigning each unit to the nearest leader. The process stops when the result stabilizes.

Two steps are performed in the generalized approach:

- the determination of the new leaders;
- the determination of the new clusters according to the new leaders.

#### 2.1.1 Determining the new leaders

Given a cluster $C$, the corresponding leader $T_C \in \mathcal{T}$ is the solution of (Eq. 1)

$$T_C = \arg \min_T \sum_{X \in C} d(X, T) = \arg \min_T \sum_{X \in C} \sum_i a_i d_i(X, T)$$

$$= \arg \min_T \sum_i a_i \sum_{X \in C} d_i(x_i, t_i) = \left[ \arg \min_{t_i} \sum_{X \in C} d_i(x_i, t_i) \right]_{i=1}^m,$$

where $t_i = [t_{i1}, \ldots, t_{ik_i}]$. By denoting $T_C = [t^*_1, t^*_2, \ldots, t^*_m]$, where $t^*_i \in (\mathbb{R}_0^+)^{k_i}$, $i = 1, 2, \ldots, m$, we obtain the following requirement: $t^*_i = \arg \min_{t_i} \sum_{X \in C} d_i(x_i, t_i)$.

Because of the additivity of the model, we can observe each variable separately and simplify the notation by omitting the index $i$.

$$t^* = \arg \min_t \sum_{X \in C} d(x, t) = \arg \min_t \sum_{X \in C} \sum_{j=1}^k w_{Xj} \delta(p_{Xj}, t_j)$$

\[\text{Springer}\]
Table 1 The basic dissimilarities and the leader \( t \), the leader \( z \) of the merged clusters and dissimilarity between merged clusters

| \( \delta(x, t) \) | \( t \) | \( z \) | \( D(C_u, C_v) \) |
|-----------------|----------|----------|----------------|
| \( \delta_1 \) | \( (p_x - t)^2 \) | \( \frac{P_t}{w_t} \) | \( \frac{w_x u + w_v v}{w_x + w_v} \) | \( \frac{w_x w_u (u - v)^2}{t} \) |
| \( \delta_2 \) | \( \frac{(p_x - t)^2}{t} \) | \( \sqrt{\frac{Q_t}{w_t}} \) | \( \sqrt{\frac{u^2 w_u + v^2 w_v}{w_x + w_v}} \) | \( \frac{w_x \frac{(u - z)^2}{z} + w_v \frac{(v - z)^2}{z}}{t} \) |
| \( \delta_3 \) | \( \frac{(p_x - t)^2}{p_x} \) | \( \frac{O_t}{P_t} \) | \( \frac{u P_x + v P_x}{P_u + P_v} \) | \( \frac{P_x \frac{(u - z)^2}{z} + P_x \frac{(v - z)^2}{v}}{P_t} \) |
| \( \delta_4 \) | \( \frac{(p_x - t)^2}{p_x} \) | \( \frac{H_t}{G_t} \) | \( \frac{H_u + H_v}{H_u + H_v} \) | \( \frac{G_u (u - z)^2 + G_v (v - z)^2}{G_t} \) |
| \( \delta_5 \) | \( \frac{(p_x - t)^2}{p_x t} \) | \( \frac{w_t}{H_t} \) | \( \frac{w_x u + w_v v}{w_x u + w_v v} \) | \( \frac{w_x \frac{(u - z)^2}{u} + w_v \frac{(v - z)^2}{v}}{t} \) |
| \( \delta_6 \) | \( \frac{(p_x - t)^2}{p_x t} \) | \( \sqrt{\frac{F_t}{H_t}} \) | \( \frac{P_x + P_x}{w_x^2 + w_v^2} \) | \( \frac{P_x \frac{(u - z)^2}{u} + P_x \frac{(v - z)^2}{v}}{w_x^2} \) |

Indices \( i \) and \( j \) are omitted.

\[
= \arg \min_t \sum_{j=1}^k \sum_{x \in C} w_{xj} \delta(p_{xj}, t_j) = \left[ \arg \min_{t_j \in \mathbb{R}^+_0} \sum_{x \in C} w_{xj} \delta(p_{xj}, t_j) \right]^{1/k}.
\]

Since the components in our model are assumed to be independent, we can optimize component-wise and omit the index \( j \).

\[
t^* = \arg \min_{t \in \mathbb{R}^+_0} \sum_{X \in C} w_X \delta(p_X, t)
\]

\( t^* \) is a kind of Fréchet mean for a selected basic dissimilarity \( \delta \).

This is a standard optimization problem with one real variable and the solution has to satisfy the condition

\[
\frac{\partial}{\partial t} \sum_{X \in C} w_X \delta(p_X, t) = 0.
\]

To simplify the notation in the following exposition, we omit superscript * and subscript \( C \), denote the component of the leader with \( t \) (before: \( t^*_i = [t_{i1}, \ldots, t_{ik_i}] \)), and add \( t \) as a subscript to the corresponding cluster.

**Leaders for \( \delta_1 \)** Here, we present the derivation for only the basic dissimilarity \( \delta_1 \) (Table 1).

The traditional clustering criterion function in the \( k \)-means and Ward’s clustering methods is based on the squared Euclidean distance dissimilarity \( d \), which is based on the basic dissimilarity \( \delta_1(p_X, t) = (p_X - t)^2 \). In our approach, the sum of weighted
squared Euclidean distance is used. For a single component, we get from (3)

\[ 0 = \sum_{X \in C} w_x \frac{\partial}{\partial t} (p_x - t)^2. \]

Therefore, the component \( t \) of the leader (optimal representative) of cluster \( C \) is

\[ t = \frac{\sum_{X \in C} w_x p_x}{\sum_{X \in C} w_x} = \frac{P_t}{w_t}. \]  

(4)

Note that \( P_t = \sum_{X \in C} w_x p_x \), where \( C \) is the cluster \( C_t \) to which the leader \( t \) belongs.

The use of the selected weights in the dissimilarity \( \delta_1 \) provides meaningful cluster representations resulting from the following two properties (subscripts \( i \) and \( j \) are reintroduced to clearly illustrate the relation):

\textbf{Property 1} Let \( w_{x_i j} = w_{x_i} \) for all \( j = 1, \ldots, k_i \). Then, for each \( i = 1, \ldots, m \):

\[ \sum_{j=1}^{k_i} t_{ij} = \frac{1}{w_t} \sum_{j=1}^{k_i} \sum_{X \in C_t} w_{x_i} p_{x_i j} = \frac{1}{w_t} \sum_{X \in C_t} w_{x_i} \sum_{j=1}^{k_i} p_{x_i j} = \frac{1}{w_t} \sum_{X \in C_t} w_{x_i} = 1 \]

In other words if the weight \( w_{x_i j} \) is the same for all components of variable \( V_i \), \( w_{x_i j} = w_{x_i} \), then for \( \delta_1 \) the leaders’ vectors \( t^*_i \) are distributions.

\textbf{Property 2} Further, let \( w_{x_i j} = n_{x_i} \) for all \( j = 1, \ldots, k_i \). Then, for each cluster \( C_t \), each variable \( V_t, i = 1, \ldots, m \), and each component \( j, j = 1, \ldots, k_i \), the following holds:

\[ t_{ij} = \frac{\sum_{X \in C_t} n_{x_i} p_{x_i j}}{\sum_{X \in C_t} n_{x_i}} = \frac{\sum_{X \in C_t} f_{x_i j}}{\sum_{X \in C_t} n_{x_i}} = f_{t_i j} = p_{t_i j}. \]

Note that in this case, the weight \( w_{x_i j} \) is constant for all components of the same variable. This result provides a solution to problem 2.

For the basic dissimilarities \( \delta \) the corresponding optimal leader \( t \), the leader \( z \) of the merged clusters and the dissimilarity \( D \) between clusters are given in Table 1. Derivations for \( \delta_k, k \in \{2, \ldots, 6\} \) are presented in the Supplement, Sect. S1.2.

\subsection*{2.1.2 Avoiding count zeros}

The dissimilarities \( \delta_4, \delta_5 \) and \( \delta_6 \) are restricted to SOs with all \( p_{x_i j} > 0 \) due to possible division with zero (this happens each time a units’ variable component equals zero). The problem of count (rounded) zeros can be resolved by the use of additive smoothing (Manning et al. 2008) known from the field of information theory. There, the empirical probability of unseen events is increased and decreased for the seen events by adding nonnegative \( \epsilon \) to each of the observed \( p_{x_i j} \)

\[ p^*_{x_i j} = \frac{p_{x_i j} + \epsilon}{\sum_{j=1}^{k_i} (p_{x_i j} + \epsilon)}. \]
This corresponds to the use of a uniform distribution as a Bayesian prior. We use/apply the Krichevsky-Trofimov (KT) smoothing estimator (Krichevsky and Trofimov 1981) to original data, where for \( \epsilon \), the ballast factor, we use fixed, equal ballast of \( 1/k_i \) for each \( p_{x_i j} \). Note that in the case when \( p_x \) are obtained from small number of micro units, additive smoothing might not be the preferred procedure. The problem of zeros is a well known preprocessing issue in the field of compositional data (Filzmoser et al. 2018, Ch. 13), where also solutions for structural zeros (which we do not encounter here) are proposed.

2.1.3 Determining new clusters

Given a list of leaders \( T \), the corresponding optimal partition \( C^* \) is determined from

\[
P(C^*) = \sum_{X \in U} \min_{T \in T} d(X, T) = \sum_{X \in U} d(X, T_{c^*(X)}),
\]

where \( c^*(X) = \arg \min_k d(X, T_k) \). The function \( c^* \) assigns each unit \( X \) to its closest leader \( T_k \in T \).

In the case that a cluster becomes empty, usually the most distant unit from another cluster is assigned to it. In the current version of R package clamix (Batagelj and Kejžar 2016), the most dissimilar unit from all the leaders from clusters with at least two units is assigned to the empty cluster.

2.2 Hierarchical method

The agglomerative hierarchical clustering procedure is a step-by-step merging of the two closest clusters starting from the clustering in which each unit forms its own cluster. The computation of dissimilarities between the new (merged) cluster and the remaining other clusters has to be specified.

To obtain the compatibility with the adapted leaders method, we define the dissimilarity between clusters \( C_u \) and \( C_v \), \( C_u \cap C_v = \emptyset \), as (Batagelj 1988)

\[
D(C_u, C_v) = p(C_u \cup C_v) - p(C_u) - p(C_v).
\]

Let \( u_i \) and \( v_i \) be \( i \)-th values of variables of leaders \( U \) and \( V \) of clusters \( C_u \) and \( C_v \), and \( z_i \) is a component of leader \( Z \) of cluster \( C_u \cup C_v \). For the basic dissimilarity \( \delta_1 \), after some computation (see Supplement, Sects. S1.1, S1.2.1), we obtain the following relation

\[
D(C_u, C_v) = \sum_i a_i \sum_j \frac{w_{uij} \cdot w_{vij}}{w_{uij} + w_{vij}} (u_{ij} - v_{ij})^2.
\]

This is a generalized Ward’s relation (with weights and with more variables). Note that this relation also holds for singletons \( C_u = \{ X \} \) or \( C_v = \{ Y \}, X, Y \in U \).

Special cases of the generalized Ward’s relation When the weights are the same for all variables \( V_i, i = 1, \ldots, m \) (i.e., \( w_{x_i} = w_x \)), we get the simplified version of the generalized Ward’s relation expressed only with cluster weights \( w_u \) and \( w_v \). In the case
when all weights for each variable \( V_i \) equal 1 (\( w_{x_i} = 1 \)), it holds that \( \sum_{X \in C} 1 = |C| \), and for each variable \( V_i, i = 1, \ldots, m \), we get the classical Ward’s relation.

### 2.3 Huygens theorem for \( \delta_1 \)

Let \( t_U \) denote the leader of the cluster consisting of all units \( U \). Then, we define (Batagelj 1988)

\[
TI = \sum_{X \in U} d(X, t_U) \quad \text{(total inertia)}
\]

\[
WI = P(C) = \sum_{C \in C} \sum_{X \in C} d(X, t_C) \quad \text{(inertia within clusters)}
\]

\[
BI = \sum_{C \in C} d(t_C, t_U) \quad \text{(inertia between clusters)}.
\]

It is easy to see that the following equation holds for \( \delta_1 \):

\[
TI = WI + BI. \quad (8)
\]

In this case, it represents the decomposition of the sum of squares and is therefore a basis for the generalization of the analysis of variance.

For each selected dissimilarity \( \delta \) and a given set of units \( U \), the value of total inertia \( TI \) is fixed. Therefore, if Huygens theorem holds, the minimization of the within-cluster inertia \( WI = P(C) \) is equivalent to the maximization of the inertia between clusters \( BI \). Huygens’ theorem is applicable for \( \delta_1 \) from Table 1. \( TI \) and \( WI \), however, can be calculated with all the dissimilarities. The \( WI/TI \) ratio can be used as a quality criterion, interpreted as the normalized gain between “partition into one cluster” and the “partition into k clusters” obtained by a clustering method.

### 3 Simulations

The purpose of the simulations is to justify the usefulness of the proposed clustering adaptations in special data settings: (1) when we have second order units of different sizes and (2) when second order units present high peaks in the empirical probability distributions of the variables. Selected synthetic data sets were constructed. We evaluated the clustering results through an external validity index that measures the agreement of the obtained and synthetically constructed partitions.

#### 3.1 Synthetic data sets

To evaluate the performance of the two proposed adaptations, (1) the inclusion of unit weights and (2) the inclusion of relative dissimilarity measure, we constructed data sets for two Monte Carlo experiments with varying important factors. Each experiment consisted of 100 generations of a data set of 100 randomly constructed units (according
to the rules described below). The results from hierarchical methods were compared according to the adjusted Rand index (ARI). The most suitable number of clusters was determined based on the CH-index (Calinski and Harabasz 1974). Additionally, we simulated experiments on 500 units where we chained the two compatible methods. We used the representatives of the best result of 50 runs of adapted leaders methods in the 10 final clusters and applied the adapted hierarchical method to these leaders. The results on 500 units were similar (comparable) to the results on 100 units with only the adapted hierarchical clustering method applied (see Supplement for the results). This was expected since the two chained methods are based on the same criterion function. The results show that for larger data sets, the chaining of the two adapted methods instead of using just adapted hierarchical method may save computation time.

The aim was to show that (1) when clustering aggregated data with the inclusion of weights, larger units are considered differently than small units. This is a result of considering each micro (first order) unit in the same way. Further, when (2) considering the relative dissimilarity measure in clustering, clusters where variables present one very highly probable category are not favored. This can result in better performance (higher ARI) in the final clustering.

To produce the modal multi-valued data, micro units were randomly generated and further aggregated into symbolic (second order) units. Symbolic units were used in the clustering process. Each micro unit consisted of two 3-category variables, where each category was selected with a certain probability. The number of micro units to aggregate was determined by a mixed binomial distribution \( \alpha B_1 + (1 - \alpha) B_2 \), where \( B_1 \sim Bin(40, 0.5) \) (for simulation of larger symbolic units) and \( B_2 \sim Bin(10, 0.1) \). \( \alpha \in [0, 1] \) enables control of the proportion of larger units. We added 1 to the final value to avoid zeros (see Supplement, Fig. S1, for a plot of the distribution). Number of clusters was 2 or 3 (depending on the experiment), which were equally probable. In the paper, we included graphical results only for the settings with a non-separable second variable, which most clearly show the improvement of adapted methods compared to the classical method. The other two possibilities are presented in greater detail in the Supplement, Sects. S2.2, S2.3.

3.2 Unit weights (experiment 1)

The use of weights was motivated by Problems 1 and 2. We would like for the symbolic units aggregated from many micro units to influence the final cluster representatives more than the rest. To determine whether this is true, we constructed 2 clusters with the first variable very separable and the second variable either (a) non-separable, (b) slightly separable or (c) separable. The category probabilities were as follows:

- Cluster 1
  - \( V_1: (0.7, 0.2, 0.1) \)
  - \( V_2: (\text{a} \ 1/3, 1/3, 1/3) \ (\text{b} \ 0.375, 0.375, 0.25) \ (\text{c} \ 0.7, 0.2, 0.1) \)

- Cluster 2
  - \( V_1: (0.1, 0.2, 0.7) \)
The distributions are presented in Fig. 1. To control for the proportion of large symbolic units, \( \alpha \) from the mixed binomial distribution was varied from 0.1 to 0.9. Since it has been shown for dissimilarity \( \delta_1 \) that the information about the variable distribution is preserved when using weights, adapted hierarchical clustering with \( \delta_1 \) was used.

The results (ARI values with 95% confidence bounds) for different \( \alpha \) for data sets with non-separable \( V_2 \) can be seen in Fig. 2. The ARI values for the weighted data set

\[ V_2: (a) \ (1/3, 1/3, 1/3) \ (b) \ (0.25, 0.375, 0.375) \ (c) \ (0.1, 0.2, 0.7) \]
are slightly better for low and medium values of $\alpha$, which corresponds to the data sets that have a non-negligible number of small symbolic units. The difference disappears for large $\alpha$, where most units are similar in size ($\approx 20$ micro units). The difference in ARI between the used methods increases slightly with a more separable variable 2 (see Supplement, Fig. S2). The use of weights diminishes the importance of the small units, which can also be seen from the dendrogram (see Supplement, Fig. S4, for an example). The clusters are more homogenous in size when using weights. For clusters without weights, one cluster increases in size because the small units’ importance alters the optimal clustering representatives. They adapt better to the small unit characteristics.

### 3.3 Relative dissimilarity measure (experiment 2)

The use of relative dissimilarity measures $\delta_k$, $k \in \{2, \ldots, 6\}$ was motivated by Problem 3. Three clusters were constructed: the first two had first variables with probability distributions with high peaks. To control for the height of the peak in cluster 1, the probability of the first category was varied from $\beta = 0.1$ to 0.9. $V_1$ in the third cluster was separable from the other two. Again (as in the first experiment), the second variable was either (a) non-separable, (b) slightly separable and (c) separable.

- **Cluster 1**
  - $V_1$: $(\beta, 1 - \beta, 0)$ where $\beta \in \{0.1, 0.2, \ldots, 0.9\}$
  - $V_2$: (a) $(1/3, 1/3, 1/3)$ (b) $(0.375, 0.375, 0.25)$ and (c) $(0.7, 0.2, 0.1)$

- **Cluster 2**
  - $V_1$: $(0, 0.9, 0.1)$
  - $V_2$: (a) $(1/3, 1/3, 1/3)$ (b) $(0.25, 0.375, 0.375)$ and (c) $(0.1, 0.7, 0.2)$

- **Cluster 3**
  - $V_1$: $(0, 0.45, 0.55)$
  - $V_2$: (a) $(1/3, 1/3, 1/3)$ (b) $(0.375, 0.25, 0.375)$ and (c) $(0.1, 0.2, 0.7)$

See Fig. 3 for a visual representation of the probability distributions for all three settings. In order not to add additional variability to the model weights were excluded from the data set (symbolic units were constructed as described with $\alpha = 1$ and only relative values were clustered). The results of hierarchical clustering with $\delta_1$ was compared to all the rest $\delta_2$ to $\delta_6$. Since the dissimilarities $\delta_4$ to $\delta_6$ require $p_x > 0$ we used KT estimator (Krichevsky and Trofimov 1981) for the empirical probabilities from the data. With the simulations (Suppl., Fig. S5) we show also that the final results with original or KT data for $\delta_1$ are very similar.

The results (ARI values with 95% confidence bounds) for different $\beta$ values and non-separable $V_2$ are displayed in Fig. 4. It can be observed that for $\beta < 0.5$ (which corresponds to the highest probability on the same category of $V_1$ for clusters 1 and 2) dissimilarities $\delta_2$, $\delta_3$ and $\delta_5$ perform better. Dissimilarities $\delta_2$ and $\delta_3$ perform best and are in the figures almost completely overlapping. However, for larger values of $\beta$, the two distributions of $V_1$ become increasingly separable and dissimilarities $\delta_1$, $\delta_2$, $\delta_3$ and $\delta_5$ give similar results. ARI values show that the dissimilarities $\delta_4$ and $\delta_6$ behave
Fig. 3 Population probability distributions for variables $V_1$ and $V_2$ for experiment 2. $V_2$ is a non-separable, b slightly and c separable.

Fig. 4 ARI values with 95% CI for the synthetic data experiment when using different dissimilarity measures for varying distribution of the first variable ($\beta$) in the first cluster. Note $\delta_2$ and $\delta_3$ are almost completely overlapping.
Fig. 5 CH-indices for different numbers of clusters with 95% CI for the synthetic data experiment when using different dissimilarity measures. The separated graphs belong to the simulated situations for different $\beta$ ($\beta = \{0.1, 0.2, 0.4, 0.8\}$). Note $\delta_2$ and $\delta_3$ are almost completely overlapping worse than expected. For these two dissimilarities the influence of an individual unit is too large. We discard these two measures from further clusterings.

To confirm this observation, the CH-index was inspected for all $\delta$s in the same situation for numbers of clusters from 2 to 6 (Fig. 5). Each graph corresponds to one $\beta$. For small values of $\beta$ (i.e., $\beta < 0.4$), $\delta_1$ as well as $\delta_5$ prefer clustering into only two (rather than three) simulated clusters. Synthetic clusters 1 and 2 are (due to the high probability on the same category at $V_1$) non-separable for that dissimilarity. The results for other simulated situations (more separability in $V_2$) are not substantially different due to the larger initial separability (see Supplement, Figs. S6–S8).
4 Applications to real-world data sets

First, the demonstration of the effect of considering weights (Problem 2) is shown for the data set of the age-sex structures (population pyramids) of the world’s countries for the year 2006. Because this data set is small, we also applied the proposed dissimilarity measures on it, excluding $\delta_4$ and $\delta_6$. Further, the behavior of the dissimilarities of $\delta_1$, $\delta_2$, $\delta_3$ and $\delta_5$ (Problem 3) is examined for the US patent citations data that could be related to certain types of clusters and characteristics (see Kejžar et al. 2011). Finally, the ESS data set, which is related to the question of Problem 1, is analyzed. The focus is on the European household structure: one variable is based on one micro unit, and the other three variables are based on the number of household members.

Due to large number of initial units, the clustering in the last two data sets was performed by chaining compatible (according to $\delta$) methods. First, the best partition (in 100 runs) into 20 clusters (a sufficiently large number to conduct further clustering) is obtained with adapted leaders method; then, a hierarchy is built on the leaders with the agglomerative clustering method.

The real-world examples do not have any known cluster structure; clusterings are done in exploratory fashion. The general background information of the real-world examples provides intuition about the correct clustering result. The sensibility of the results is also evaluated based on the proportion of the within-cluster inertia to the total inertia (i.e., the normalized gain, see Sect. 2.3).

4.1 Age-sex structures of the world’s countries for 2006

The data set consists of 223 countries, each described by two symbolic variables: the age structure for males and females according to 5-year age groups obtained from IDB (2008). Since the number of units is relatively small, hierarchical clustering was performed. In this special data set, the interpretations are most appropriate/meaningful if the weighted and unweighted cases of $\delta_1$ are used. In the weighted case, the cluster representatives are the age-sex structures of the whole included population (from all countries included in the cluster), whereas in the unweighted case, only the shapes of the structures relative to each sex are considered (each country counts as one unit to cluster). The results with the included weights provide a more straightforward interpretation.

In our approach, the dissimilarity between two clusters in the agglomerative hierarchical clustering is defined as the difference between the error of the joint cluster minus the errors of both clusters. Since the criterion function is defined as a sum of cluster errors, the dissimilarity between two clusters that join at each step is equal to the change in the criterion function. Therefore, we use these values as a criterion to determine the most natural number of clusters, i.e., where to cut the dendrogram to obtain the most natural partition. The obtained partitions are compared with contingency tables and with the ARI.

The dendrogram obtained with agglomerative hierarchical clustering using $\delta_1$ is presented in Fig. 6 for the unweighted case and in Fig. 7 for the weighted case. To provide a closer look at the lower part of the dendrogram (with many clusters), we
Fig. 6 Dendrogram for $\delta_1$, without weights (distributions relative by sex)
Fig. 7  Dendrogram for $\delta_1$ with male and female population size included
Table 2  Contingency table comparing the partitions into six clusters obtained with the same dissimilarity $\delta_1$ with and without weights (with and without male/female population size)

| $\delta_1$ | Weights | Total |
|------------|---------|-------|
|            | 1       | 2     | 3     | 4     | 5     | 6     |
| No weights | 1       | 45    | 0     | 0     | 0     | 0     | 45    |
|            | 2       | 16    | 29    | 0     | 3     | 0     | 48    |
|            | 3       | 0     | 0     | 0     | 0     | 0     | 3     |
|            | 4       | 0     | 0     | 15    | 26    | 0     | 1     | 42    |
|            | 5       | 0     | 0     | 0     | 0     | 47    | 0     | 47    |
|            | 6       | 0     | 0     | 4     | 0     | 2     | 32    | 38    |
| Total      | 61      | 29    | 19    | 29    | 49    | 36    |

plot the square root of the dendrogram heights (here the dissimilarities among joint clusters). This decreases the large heights and suggests a natural partition into six clusters. The original dendrograms are presented in the Supplement, Fig. S12. The cluster representatives are presented above each of the six clusters. They represent the averages of the components in the unweighted case and the weighted averages in the weighted case. In the weighted case only, they represent the real age-sex distributions of the population. In the unweighted case, the third cluster contains only three countries (Kuwait, UAE, and Northern Mariana Islands (NMI), see also Table 2) with very asymmetrical shapes. They exhibit a large number of 20- to 35-year-olds (males, presumably workers in the construction and service sectors, for the Gulf states; females for NMI). Because of the small cluster size, its representative is presented with a different scale (see also Supplement). When population size is included in the clustering process, this asymmetric cluster is further joined with Saudi Arabia and is recognized as an inhomogeneous subcluster (the first subcluster of the sixth cluster in the dendrogram in Fig. 7). As in the weighted case, Saudi Arabia is included in the last of the six clusters in the unweighted case. The population pyramids for the four mentioned countries are presented in the Supplement, Fig. S9.

Weighted clustering clearly detects one cluster with an age-sex distribution with more bulges and notches. This is the second cluster in Fig. 7, which characterizes the age-sex distribution of the population of China. By contrast, this cluster is not noticeable in the partition into 6 clusters obtained without weights (see Fig. 6). In the weighted case, China forms a separate subcluster in the second of six clusters, and because of its size, it has a strong influence on the cluster representative. China is also in the second cluster in the unweighted case, where its closest country is Cuba. Similarly, the closest country to Japan is Monaco in the unweighted case (see Supplement, Figs. S10, S11), whereas they are not closely connected in the weighted case.

Additional comparisons of both partitions are made with respect to the number of common countries in both clusterings in Table 2. ARI for the partitioning into 6 clusters obtained with and without weights is 0.696.

In addition to the $\delta_1$ comparison all the other combinations (except $\delta_4$ and $\delta_6$) for clustering were run, and the overall conclusion is that all the dissimilarity-weight combinations produce the expected results (i.e., four final clusters) associated with
the main age-sex structure distributions from the demographic literature (the expanding, stationary and contracting types of population pyramids). As expected, the strong influence of China on its cluster representative is noted in all the weighted clustering results. The dendrograms and ARI among the clusterings and their cluster representatives can be found in the Supplement, Sect. S3.2.

4.2 United States patents data

US patents data were downloaded from NBER (2016). The data set is explained in detail in Hall et al. (2001) and on the website: only the relevant data are mentioned here. We analyzed patents regarding two symbolic variables: number of citations received and technological category of the citing patents. The first variable number of citations is presented with a distribution over years in the observed 20-year period from 1987 to 2006. The second variable technological category is presented with a distribution over six broad areas of technology [determined by Hall et al. (2001)]: Chemical, Computer&Communication (CC), Drugs&Medical (DM), Electrical&Electronics (EE), Mechanical, and Others. Missing values (only 9) are considered as a separate value and are denoted by MV. The analysis is limited to patents with at least 20 citations; there are 12,291 such patents.

We observe the clustering results produced with dissimilarities $\delta_1, \delta_2$ (and $\delta_3, \delta_5$ in the Suppl., Sect. S4.2), either considering relative (empirical probability) distributions, i.e., the unweighted case or with the total number of citations included as weights for both variables ($w_{x1} = w_{x2} = n_x$).

In the unweighted clustering, $\delta_1$ clearly identifies six clusters which are very well characterized by the technological categories (Fig. 8). All the technological variable distributions have only one evident peak (highly probable category) at the mode value, which confirms the observations described in Kejžar et al. (2011) that dissimilarity $\delta_1$ favors the largest value of a variable. $\delta_2$ on the other hand, identifies three main clusters, where most similar categories are merged together (i.e., cluster 1: CC and EE, cluster 3: Chemical and DM and cluster 2: the rest), see Fig. 9. Clearly, the technological category is the main separator variable.

The other variable, number of citations, indicates differences in citation culture among technological categories. In the fast developing areas of CC and EE, the largest number of citations appear early (1989), after which we see a decreasing trend over the observed period. However, for Chemical and DM, the largest number of citations appears, approximately, in the year 1998. For the second cluster (see Fig. 9), the number of citations is rather evenly distributed over time.

For relative dissimilarities $\delta_3$ and $\delta_5$ (see Suppl.) 2 and 4 main clusters are determined, respectively. The results (representatives) differ substantially from results from $\delta_1$ and $\delta_2$ due to the different nature of dissimilarity. The aim of not favoring the largest value of a variable is less pronounced there.

A more detailed overview over all eight clustering possibilities can be found in the Supplement. The main finding is that the influence of weighting in the case of $\delta_1$ is almost negligible in this data set, and the results with all relative dissimilarity measures differ noticeably from the unweighted results. We still identify the same
Fig. 8 Dendrogram and cluster representatives for $\delta_1$ (6 final clusters)
number of main clusters in both cases, with and without weights. In the case of $\delta_2$ the cluster of dominant Chemical and DM is not as technologically homogenous as in the unweighted case, whereas the cluster with prevailing mechanical and others is more homogenous.

4.3 ESS data

The data set ESS (2010) is an output from an academically driven social survey. Its main purpose is to gain insight into the behavior patterns, beliefs and attitudes of European populations (ESS website 2012). We used data from Round 5 (conducted in 2010), which included more than 50,000 respondents. The focus was on the variables that describe household structure: (1) the genders of people in the household, (2) the relationships between the respondent and the members of the household (3) the years of birth of the people in household and (4) the country of residence of respondent, i.e., the country in which the household is located. The respondent answered the first three
questions for every member of his/her household. Symbolic variables (with counts of household members) were constructed from these variables. Variable 3 (age of household members) is a numeric variable, and a decision about the category borders was required. Here, we present the clustering results with the categorization of the age variable into ten-year groups. Additional categorization (into economic groups of the working population) is described in the Supplement, Sect. S5.1. The variables were:

- \( V_1 \): gender (2 components):
  \( \text{male} : f_{11}, \text{female} : f_{12} \)

- \( V_2 \): categories of household members (7 components, respondent constantly 1):
  \( \text{respondent} : f_{21} = 1, \text{partner} : f_{22}, \text{offspring} : f_{23}, \text{parents} : f_{24}, \text{siblings} : f_{25}, \text{relatives} : f_{26}, \text{others} : f_{27} \)

- \( V_3 \): year of birth for every household member transformed into age (10 components):
  \( \{ 0 – 9, 10 – 19, 20 – 29, 30 – 39, 40 – 49, 50 – 59, 60 – 69, 70 – 79, 80+, \text{MV} \} \)

- \( V_4 \): country of residence (26 components, all but one with value zero):
  \( \{ \text{Belgium} : f_{4,1}, \ldots, \text{Ukraine} : f_{4,26} \} \)

There were 641 respondents with missing values for year of birth; it seemed reasonable to add the category MV to variable \( V_3 \). This method of handling missing values is naive and could possibly lead to biased results (i.e., it is possible that the birth years of very old or non-related family members are most likely to be missing, so they could form a special pattern). A refined clustering analysis would use an alternative well-known imputation method [e.g., multiple imputation, Rubin (1987)]. Note that for each unit (respondent) in the data set, the components of variables \( V_1 \) to \( V_3 \) sum to a constant number (the number of all household members of that respondent). For the last variable \( V_4 \), the sum equals 1.

Design and population weights are supplied by the data set for each unit (respondent). To obtain results that are representative of the EU population, both weights should also be used for households. Because special weights for households are not available, we used the weights provided in the data set in our demonstration: before clustering, each unit’s (respondent’s) symbolic variables were multiplied by the design and population weights. \( w_{V_i} \) (used in the clustering process) for variables \( V_1 \) to \( V_3 \) is equal to the number of household members multiplied by both supplied weights, and for \( V_4 \), it is the product of both weights alone.

To gain an overview of the main European household patterns two step clustering (chaining the adapted leaders and agglomerative method based on dissimilarity \( \delta_1 \)) was used to perform clustering of the 50,372 units (households). The natural choice was to use clustering with weights included. The results of clustering without weights (which also discards the design and population weights) are described in the Supplement, Sect. S5.2. The number of final clusters was selected by eyeballing the dendrogram and selecting the cutpoint as the location where the dissimilarity among clusters had the highest jump (apart from clustering into only two groups).

The representatives (variable distributions) of the final clustering are presented in Figs. 10 and 11. The best clustering splits the data into one very large cluster (\( C_4 \) with
Fig. 10 Representatives for the final 4 clusters (four variables included, first three presented)
28,912 households), one medium sized cluster ($C_2$ with 9595 households) and two small clusters ($C_1$ with 5507 and $C_3$ with 6358 households).

Immediately, we can note that cluster $C_1$ is dominated by extended families in the Russian Federation. This cluster is also the smallest. The largest cluster, $C_4$, has a much more even distribution of households over countries. The shares of core families and younger respondent in the second smallest cluster, $C_3$, are highest in Israel, Slovenia, Czech Republic, Poland and Croatia. In these countries, offspring stay with their parents a long time before becoming independent.

The country of residence (as a highly influential variable) was excluded from the clusterings where comparison between two age categorizations was performed (Supplement). The final four clusters show similar patterns, although a more refined categorization reveals more information. In the clustering results with no weights
included, small households gain influence. This is reflected by the cluster representa-
tives.

5 Conclusion

Versions of well-known leaders nonhierarchical and agglomerative hierarchical meth-
ods, adapted for modal-valued symbolic data, are presented in this paper. Their
selection was motivated by real-world data problems, such as clustering regions based
on demographic structures and clustering works based on their citations and their
other characteristics. Since data measured on traditional measurement scales (numer-
aical, ordinal, categorical) can be transformed into modal symbolic representations the
methods can be used to cluster data sets of mixed units. Our approach enables users to
consider the original frequency information by using weights. The proposed clustering
methods are compatible—they solve essentially the same optimization problem and
can be used separately or in combination (usually for large data sets). The optimization
criterion function depends on a basic dissimilarity $\delta$, which enables the specification
of different criteria. In principle, we could use different $\delta$s for different symbolic
variables because of the additivity of the components of the criterion function.

The difference in our approach relative to other clustering approaches for modal-
valued data is that it is focused on the interpretability of the final cluster representatives.
The inclusion of weights allows the cluster leaders to preserve the structure (i.e., real
sex-age structure in the clustering of population pyramids). Additionally, the use of a
different (relative) dissimilarity measure between a unit and a cluster representative
prevents the clustering process from favoring only one (mode) value of a variable and
therefore produces more informative optimal cluster representatives (i.e., US patent
citation patterns and their technological categories).

The usefulness of the proposed methods was assessed by evaluating the clustering of
carefully selected artificially constructed data sets. The methods were applied to three
different real-world data sets: age-sex structures of world countries, US patents 1987–
2006 and household structures from the ESS 2010. The clusters of age-sex structures
gain in interpretability the most when using weighted clustering with dissimilarity
$\delta_1$. Other dissimilarity-weight combinations, except $\delta_4, \delta_6$, also produce the expected
results (i.e., the four final clusters of age-sex structure distributions from demographic
literature). The cluster representatives of the United States patent citation patterns are
more informative when $\delta_2$ is used since the technological category is not favored.
In the data set of household structures from the ESS 2010, the variables are based
on different numbers of micro units. This is taken into account when clustering with
weights ($\delta_1$).

The proposed approach is implemented in the R package clamix (Batagelj and
Kejžar 2016).

References

Anderberg MR (1973) Cluster analysis for applications. Academic Press, New York
Batagelj V (1988) Generalized ward and related clustering problems. In: Bock HH (ed) Classification and related methods of data analysis. North-Holland, Amsterdam, pp 67–74
Batagelj V, Kejžar N (2016) Clamix—clustering symbolic objects. Program in R. https://r-forge.r-project.org/projects/clamix/. Accessed 21 Oct 2020
Billard L, Diday E (2003) From the statistics of data to the statistics of knowledge: symbolic data analysis. J Am Stat Assoc 98(462):470–487
Billard L, Diday E (2006) Symbolic data analysis. Conceptual statistics and data mining. Wiley, Chichester
Billard L, Diday E (2019) Clustering methodology for symbolic data. Wiley, New York
Bock HH, Diday E (eds) (2000) Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data. Springer, Heidelberg
Calinski T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat Theory Methods 3(1):1–27
de Carvalho FAT, Sousa RMCR (2010) Unsupervised pattern recognition models for mixed feature-type symbolic data. Pattern Recogn Lett 31:430–443
de Carvalho FAT, Brito P, Bock HH (2006) Dynamic clustering for interval data based on L2 distance. Comput Stat 21(2):231–250
Diday E (1979) Optimisation en classification automatique. Tome 1.. INRIA, Rocquencourt (in French) Diday E, Noirhomme-Fraiture M (2008) Symbolic data analysis and the SODAS software. Wiley, Chichester
ESS Round 5: European Social Survey Round 5 Data (2010). Data file edition 2.0. Norwegian Social Science Data Services, Norway—Data Archive and distributor of ESS data
ESS website. http://www.europeansocialsurvey.org/. Accessed 27 Sept 2012
Filzmoser P, Hron K, Templ M (2018) Applied compositional data analysis. Springer, Switzerland
Gowda KC, Diday E (1991) Symbolic clustering using a new dissimilarity measure. Pattern Recogn 24(6):567–678
Hall BH, Jaffe AB, Tratjenberg M (2001) The NBER patent citation data file: lessons, insights and methodological tools. NBER Working Paper 8498, NBER, 2001
Hardy A, Lallemant P (2004) Clustering of symbolic objects described by multi-valued and modal variables.
In: Banks D, House L, McMorris F, Arabie P, Gaul W (eds) Classification, clustering and data mining applications. Springer, Berlin, pp 325–332
Hartigan JA (1975) Clustering algorithms. Wiley, New York
Ichino M, Yaguchi H (1994) Generalized Minkowski metrics for mixed feature type data analysis. IEEE Trans Syst Man Cybern 24(4):698–708
IDB: International Data Base (2008). http://www.census.gov/ipc/www/idbnew.html. Accessed 10 Feb 2008
Irpino A, Verde R (2006) A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In: Batagelj V, Bock HH, Ferligoj A, Ziberna A (eds) Data science and classification. Springer, Berlin, pp 185–192
Irpino A, Verde R, de Carvalho FAT (2014) Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. Expert Syst Appl 41:3351–3366
Kejžar N, Korenjak-Černe S, Batagelj V (2011) Clustering of distributions: a case of patent citations. J Classif 28(2):156–183
Kim J, Billard L (2011) A polythetic clustering process and cluster validity indexes for histogram-valued objects. Comput Stat Data Anal 55:2250–2262
Kim J, Billard L (2012) Dissimilarity measures and divisive clustering for symbolic multimodal-valued data. Comput Stat Data Anal 56(9):2795–2808
Kim J, Billard L (2013) Dissimilarity measures for histogram-valued observations. Commun Stat Theory Methods 42(2):283–303
Korenjak-Černe S, Batagelj V (1998) Clustering large data sets of mixed units. In: Rizzi A, Vichi M, Bock HH (eds) 6th Conference of the international federation of classification societies (IFCS-98) Universita “La Sapienza”, Rome, 21–24 July 1998. Advances in data science and classification. Springer, Berlin, pp 43–48
Korenjak-Černe S, Batagelj V (2002) Symbolic data analysis approach to clustering large datasets. In: Jajuga K, Sokolowski A, Bock HH (eds) 8th Conference of the international federation of classification societies, 16–19 July 2002, Cracow, Classification, clustering and data analysis. Springer, Berlin, pp 319–327
Korenjak-Černe S, Batagelj V, Japelj Pavešič B (2011) Clustering large data sets described with discrete distributions and its application on TIMSS data set. Stat Anal Data Min 4(2):199–215
Korenjak-Černe S, Kejžar N, Batagelj V (2015) A weighted clustering of population pyramids for the world’s countries, 1996, 2001, 2006. Popul Stud 69(1):105–120
Košmelj K, Billard L (2011) Clustering of population pyramids using Mallows’ L2 distance. Metodološki zvezki 8(1):1–15
Krichevsky RE, Trofimov VK (1981) The performance of universal encoding. IEEE Trans Inf Theory IT–27(2):199–207
Manning C, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge
NBER Patent Data Project. https://sites.google.com/site/patentdataproject/Home. Accessed 6 Sept 2016
Noirhomme-Fraiture M, Brito P (2011) Far beyond the classical data models: symbolic data analysis. Stat Anal Data Min 4(2):157–170
Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York
Verde R, Irpino A (2010) Ordinary least squares for histogram data based on Wasserstein distance. In: Proc. COMPSTAT’2010. Springer, Berlin, pp 581–589
Verde R, de Carvalho FAT, Lechevalier Y (2000) A dynamic clustering algorithm for multi-nominal data. In: Kiers HAL, Rasson JP, Groenen PJF, Schader M (eds) Data analysis, classification, and related methods. Springer, Berlin
Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Nataša Kejžar1 · Simona Korenjak-Černe2 · Vladimir Batagelj3,4,5

Nataša Kejžar
natasa.kejzar@mf.uni-lj.si

Simona Korenjak-Černe
simona.cerne@ef.uni-lj.si

Vladimir Batagelj
vladimir.batagelj@fmf.uni-lj.si

1 Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Vrazov trg 2, 1000 Ljubljana, Slovenia
2 School of Economics and Business (SEB), University of Ljubljana, Ljubljana, Slovenia
3 Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia
4 Andrej Marušič Institute, University of Primorska, Koper, Slovenia
5 National Research University Higher School of Economics, Moscow, Russia