Symbolic data analysis
Clustering of large data sets of mixed units

Vladimir Batagelj

IMFM Ljubljana and IAM UP Koper

CRoNoS meeting
16 - 17 April 2016, Colegio de España, Paris
Symbolic data

Data table

<table>
<thead>
<tr>
<th></th>
<th>⋯</th>
<th>variable&lt;sub&gt;j&lt;/sub&gt;</th>
<th>⋯</th>
</tr>
</thead>
<tbody>
<tr>
<td>⋯</td>
<td>⋯</td>
<td>⋯</td>
<td>⋯</td>
</tr>
<tr>
<td>unit&lt;sub&gt;i&lt;/sub&gt;</td>
<td>⋯</td>
<td>value&lt;sub&gt;i,j&lt;/sub&gt;</td>
<td>⋯</td>
</tr>
<tr>
<td>⋯</td>
<td>⋯</td>
<td>⋯</td>
<td>⋯</td>
</tr>
</tbody>
</table>

In classical data analysis value<sub>i,j</sub> is a single element (number or label) measured in standard measurement scales (absolute, ratio, interval, ordinal, nominal).

In *symbolic data table* value<sub>i,j</sub> can be also a complex data such as: interval, set of values, distribution (in general sense), time series, tree, text, function, etc. Rules linking variables and taxonomies of values can be specified.
Symbolic data analysis aims to extend existing data analysis methods to symbolic data and to develop new ones.

It was introduced in 1987 by Edwin Diday [Diday, E. (1987)].

Three European projects:

- SODAS - Symbolic Official Data Analysis System (1996-99),
- ISO-3D - Interpretation of Symbolic Objects with 3D Representation (1998-01),

resulted in a program for symbolic data analysis SODAS 2.


Additional two books are to appear soon.

Three packages for SDA are available in R:

- RSDA,
- symbolicDA,
- Clamix

Symbolic data analysis group at LinkedIn
Aggregating data into symbolic data preserves much more information than the standard approach using mean values.

Let $\Sigma(S, V)$ denote a summary – a symbolic value of variable $V$ over the subset of units $S$.

A good summary satisfies the condition: for $S_1 \cap S_2 = \emptyset$ it holds

$$\Sigma(S_1 \cup S_2, V) = f(\Sigma(S_1, V), \Sigma(S_2, V))$$

With my collaborators (Simona Korenjak-Černe and Nataša Kejžar) we are developing the clustering algorithms for symbolic objects described by modal valued symbolic data

Clustering symbolic data

For clustering of SOs we adapted two classical clustering methods:

- *leaders method* (a generalization of k-means method [Hartigan, J. A. (1975)], dynamic clouds [Diday, E. (1979)]).

Both adapted methods are based on the same criterion function—they are solving the same clustering problem.

With the leaders method the size of the sets of units is reduced to a manageable number of leaders.

The obtained leaders can be further clustered with the compatible agglomerative hierarchical clustering method to reveal relations among them and using the dendrogram also to decide upon the right number of clusters.
Symbolic objects described with distributions

An SO $X$ is described by a list $X = [x_i]$ of descriptions of variables $V_i$. The values NA (not available) are treated as an additional category for each variable. In our model, each variable is described with frequency distribution (bar chart) of its values

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

With

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

we denote the corresponding probability distribution.

$$\sum_{j=1}^{k_i} p_{x_{ij}} = 1, \quad i = 1, \ldots, m$$
We approach the clustering problem as an optimization problem over the set of *feasible* clusterings $\Phi_k$ – partitions of units into $k$ clusters. The *criterion function* has the following form

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

The *total error* $P(C)$ of the clustering $C$ is a sum of *cluster errors* $p(C)$. 
The cluster error

There are many possibilities how to express the cluster error \( p(C) \). In this paper we shall assume a model in which the error of a cluster is a sum of differences of its units from the cluster’s representative \( T \)

\[
p(C, T) = \sum_{X \in C} d(X, T). \tag{2}
\]

Note that in general the representative needs not to be from the same ”space” (set) as units.
Representatives

The best representative is called a *leader*

\[ T_C = \arg\min_T p(C, T). \] (3)

Then we define

\[ p(C) = p(C, T_C) = \min_T \sum_{X \in C} d(X, T). \] (4)

The SO \( X \) is described by a list \( X = [x_i] \). Assume that also representatives are described in the same way \( T = [t_i] \),
\( t_i = [t_{i1}, t_{i2}, \ldots, t_{ik_i}] \).
Dissimilarity between SOs

We introduce a dissimilarity measure between SOs with

\[ d(X, T) = \sum_{i} \alpha_i d(x_i, t_i), \quad \alpha_i \geq 0, \quad \sum_{i} \alpha_i = 1, \quad (5) \]

where

\[ d(x_i, t_i) = \sum_{j=1}^{k_i} w_{xij} \delta(p_{xij}, t_{ij}), \quad w_{xij} \geq 0. \quad (6) \]

This is a kind of a generalization of the squared Euclidean distance.

The weight \( w_{xij} \) can be for the same unit \( X \) different for each variable \( V_i \) (needed in descriptions of ego-centric networks, population pyramids, etc.).
Leaders method is a generalization of a popular nonhierarchical clustering k-means method. The idea is to get "optimal" clustering into a pre-specified number of clusters with the following iterative procedure:

1. Determine an initial clustering
2. Repeat
   - Determine leaders of the clusters in the current clustering;
   - Assign each unit to the nearest new leader – producing a new clustering
3. Until the leaders stabilize.
Selection of the new leaders

Given a cluster \( C \), the corresponding leader \( T_C \) is the solution of the problem

\[
T_C = \arg\min_T \sum_{X \in C} d(X, T) = \left[ \arg\min_{t_i} \sum_{X \in C} d(x_i, t_i) \right]_{i=1}^m
\]

Therefore \( T_C = [t_i^*] \) and \( t_i^* = \arg\min_{t_i} \sum_{X \in C} d(x_i, t_i) \). To simplify the notation we omit the index \( i \).

\[
t^* = \arg\min_t \sum_{X \in C} d(x, t) = \left[ \arg\min_{t_j \in \mathbb{R}} \sum_{X \in C} w_{Xj} \delta(p_{Xj}, t_j) \right]_{j=1}^k
\]
Again we omit the index $j$

$$t^* = \arg\min_{t \in \mathbb{R}} \sum_{X \in C} w_X \delta(p_x, t)$$

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

$$\frac{\partial}{\partial t} \sum_{X \in C} w_X \delta(p_x, t) = 0$$

or

$$\sum_{X \in C} w_X \frac{\partial \delta(p_x, t)}{\partial t} = 0 \quad (7)$$
Dissimilarities $\delta$

<table>
<thead>
<tr>
<th>$\delta(x, t)$</th>
<th>$t^*_{ij}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(p_x - t)^2$</td>
<td>$\frac{P_{ij}}{A_{ij}}$</td>
</tr>
<tr>
<td>$(\frac{p_x - t}{t})^2$</td>
<td>$\frac{Q_{ij}}{P_{ij}}$</td>
</tr>
<tr>
<td>$(\frac{p_x - t}{t})^2$</td>
<td>$\sqrt{\frac{Q_{ij}}{A_{ij}}}$</td>
</tr>
<tr>
<td>$(\frac{p_x - t}{p_x})^2$</td>
<td>$\frac{H_{ij}}{F_{ij}}$</td>
</tr>
<tr>
<td>$(\frac{p_x - t}{p_x})^2$</td>
<td>$\frac{A_{ij}}{H_{ij}}$</td>
</tr>
<tr>
<td>$(\frac{p_x - t}{p_x})^2$</td>
<td>$\sqrt{\frac{P_{ij}}{H_{ij}}}$</td>
</tr>
</tbody>
</table>

$$A_{ij} = \sum_{X \in C} w_{xij}$$
$$P_{ij} = \sum_{X \in C} w_{xij} p_{xij}$$
$$Q_{ij} = \sum_{X \in C} w_{xij} p_{xij}^2$$
$$H_{ij} = \sum_{X \in C} \frac{w_{xij}}{p_{xij}}$$
$$F_{ij} = \sum_{X \in C} \frac{w_{xij}}{p_{xij}^2}$$
For $\delta_1(p_x, t) = (p_x - t)^2$ we get from (8)

$$0 = \sum_{x \in C} w_x \frac{\partial}{\partial t} (p_x - t)^2 = -2 \sum_{x \in C} w_x (p_x - t)$$

Therefore

$$t^* = \frac{\sum_{x \in C} w_x p_x}{\sum_{x \in C} w_x} = \frac{P}{A}$$
Leaders for $\delta_1$

Let $w_{xij} = w_{xi}$ then for each $i = 1, \ldots, m$:

$$\sum_{j=1}^{k_i} t_{ij}^* = \frac{1}{A_i} \sum_{j=1}^{k_i} \sum_{X \in C} w_{xi} p_{xij} = 1$$

The leaders’ components are *distributions*.

Let further $w_{xij} = n_{xi}$ then for each $i = 1, \ldots, m$:

$$t_{Cij}^* = \frac{\sum_{X \in C} n_{xi} p_{xij}}{\sum_{X \in C} n_{xi}} = \frac{\sum_{X \in C} f_{xij}}{\sum_{X \in C} n_{xi}} = \frac{f_{Cij}}{n_{Ci}} = p_{Cij}$$

The leader of a cluster is its *distribution*.
Determining the new clustering

Given leaders \( \mathbf{T} \) the corresponding optimal clustering \( \mathbf{C}^* \) is determined from

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

(8)

where

\[
c^*(X) = \arg\min_k d(X, T_k)
\]

We assign each unit \( X \) to the closest leader \( T_k \in \mathbf{T} \).
Hierarchical agglomerative clustering

The hierarchical agglomerative clustering procedure is based on a step-by-step merging of the two closest clusters.

- Each unit forms a cluster: $C_n = \{\{X\}: X \in \mathcal{U}\}$;
- They are at level 0: $h(\{X\}) = 0, \ X \in \mathcal{U}$;

**for** $k = n - 1$ **to** 1 **do**

- Determine the closest pair of clusters $(u, v) = \arg\min_{i,j: i \neq j} \{D(C_i, C_j): C_i, C_j \in C_{k+1}\}$;
- Join the closest pair of clusters $C_{(uv)} = C_u \cup C_v$
- $C_k = (C_{k+1} \setminus \{C_u, C_v\}) \cup \{C_{(uv)}\}$
- $h(C_{(uv)}) = D(C_u, C_v)$
- Determine the dissimilarities $D(C_{(uv)}, C_s), C_s \in C_k$

**endfor**

$C_k$ is a partition of the finite set of units $\mathcal{U}$ into $k$ clusters. The level $h(C)$ of the cluster $C_{(uv)} = C_u \cup C_v$. 

V. Batagelj

Symbolic data analysis
Clustering and optimization
Leaders method
Agglomerative method
Examples
References
Dissimilarity between clusters

Therefore the computation of dissimilarities between new (merged) cluster and the rest 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

$$D(C_u, C_v) = p(C_u \cup C_v) - p(C_u) - p(C_v)$$

$$= \sum_i \alpha_i \sum_j \frac{A_{uij} \cdot A_{vij}}{A_{uij} + A_{vij}} (u_{ij} - v_{ij})^2$$

(9)

a generalized Ward’s relation. $u_i$ and $v_i$ are components of the leaders of clusters $C_u$ and $C_v$. 
Instead of the squared Euclidean distance other dissimilarity measures $\delta(x, t)$ can be used (see [Kejžar, N. et al. (2011)]). Relations similar to Ward’s can be derived for them.

The proposed approach is implemented in the R-package **Clamix**.

It was successfully applied on different data sets (population pyramids, TIMSS, cars, foods, citation patterns of patents, and others).
Sheme of analysis

raw data
⇓
ENCODE
⇓
unified data
⇓
MAKE SOs
⇓
SOs - lists of distributions
⇓
leaderSO
⇓
clustering and cluster leaders
⇓
hclustSO
⇓
hierarchy and cluster leaders
⇓
ANALYSIS
⇓
dendrogram, reports
The fourth main cluster has typical expansive pyramid's shape. It includes 69 less developed (mostly African) countries. It corresponds to the Stage 1 of the DTM.

Shapes of four main clusters look very similar to these from the year 1996, although more precise observation reveals differences, including a 5-year lag toward the future.

This cluster is the most uniform among four main clusters.

The hierarchy for the year 2001 obtained on the same 215 countries as in the year 1996, though more precise observation reveals differences, including a 5-year lag toward the future.
World 2001 / 4 clusters on the map

[Korenjak-Černe, S. et al. (2015)]
### Clusters of the US counties

We used the proposed clustering method to group the US counties into clusters. From the hierarchy obtained, we present some examples of the clusters. The population pyramids for each cluster are shown below.

#### Example Clusters

- **Montmorency County, Michigan**
- **Roscommon County, Michigan**
- **Pinellas County, Florida**
- **Palm Beach County, Florida**
- **Ocean County, New Jersey**
- **Clifton Forge city, Virginia**
- **Bedford city, Virginia**
- **Martinsville city, Virginia**
- **Brunswick County, North Carolina**
- **Cherokee County, North Carolina**
- **Rabun County, Georgia**
- **Fannin County, Georgia**
- **Pickett County, Tennessee**
- **Cedar County, Missouri**
- **Kimble County, Texas**
- **Pacific County, Washington**
- **San Juan County, Washington**
- **Idaho County, Idaho**
- **Jeff Davis County, Texas**
- **Judith Basin County, Montana**
- **Roberts County, Texas**
- **Johnson County, Wyoming**
- **Cambria County, Pennsylvania**
- **Ohio County, West Virginia**
- **Oxford County, Maine**
- **St. Johns County, Florida**
- **Sullivan County, New Hampshire**
- **Belknap County, New Hampshire**
- **Madison County, Virginia**
- **Delta County, Michigan**
- **Menominee County, Michigan**
- **Kerr County, Texas**
- **Harrison County, Missouri**
- **Hyde County, South Dakota**
- **Taylor County, Iowa**
- **Holt County, Missouri**
- **Lavaca County, Texas**
- **Chowan County, North Carolina**
- **Douglas County, Minnesota**
- **Goshen County, Wyoming**
- **Douglas County, Oregon**
- **Delaware County, New York**
- **Las Animas County, Colorado**
- **Wheatland County, Montana**
- **Texas County, Missouri**
- **Goliad County, Texas**
- **Allamakee County, Iowa**
- **Van Buren County, Iowa**
- **Greene County, Iowa**
- **Polk County, Nebraska**
- **Williamson County, Tennessee**
- **St. Tammany Parish, Louisiana**
- **Rockdale County, Georgia**
- **Canadian County, Oklahoma**
- **Clinton County, Michigan**
- **Kodiak Island Borough, Alaska**
- **Teton County, Idaho**
- **Nassau County, New York**
- **Hartford County, Connecticut**
- **Essex County, Massachusetts**
- **Sonoma County, California**
- **Kitsap County, Washington**
- **Porter County, Indiana**
- **Eaton County, Michigan**
- **Dinwiddie County, Virginia**
- **Louisa County, Virginia**
- **Woodford County, Illinois**
- **Rio Grande County, Colorado**
- **Austin County, Texas**
- **Monroe County, Illinois**
- **Benton County, Iowa**
- **Kingfisher County, Oklahoma**
- **Lafayette County, Wisconsin**
- **Newton County, Indiana**
- **Tuscola County, Michigan**
- **Owsley County, Kentucky**
- **Amite County, Mississippi**
- **Panola County, Texas**
- **Etowah County, Alabama**
- **Fayette County, Indiana**
- **Bladen County, North Carolina**
- **Hocking County, Ohio**
- **Van Buren County, Tennessee**
- **Wayne County, West Virginia**
- **Blount County, Tennessee**
- **Sevier County, Tennessee**
- **Cocke County, Tennessee**
- **Marion County, Tennessee**
- **Hancock County, Mississippi**
- **Stewart County, Tennessee**
- **Hampshire County, West Virginia**
- **Glascock County, Georgia**
- **Massac County, Illinois**
- **Barbour County, West Virginia**
- **Mineral County, West Virginia**
- **Yates County, New York**
- **Latimer County, Oklahoma**
- **Madison County, Nebraska**
- **Codington County, South Dakota**
- **Seneca County, Ohio**
- **Erie County, Pennsylvania**
- **Menifee County, Kentucky**
- **Washington County, Georgia**
- **St. Clair County, Illinois**
- **Baker County, Georgia**
- **Grady County, Georgia**
- **Huron County, Ohio**
- **Perry County, Ohio**
- **Camden County, New Jersey**
- **Winnebago County, Illinois**
- **Washington County, Indiana**
- **Marshall County, Tennessee**
- **Benton County, Oregon**
- **Ionia County, Michigan**
- **Prince George County, Virginia**
- **Newport News city, Virginia**
- **Clark County, Nevada**
- **Orange County, California**
- **Barrow County, Georgia**
- **Johnston County, North Carolina**
- **Bartow County, Georgia**
- **DeSoto County, Mississippi**
- **Hoke County, North Carolina**
- **Montgomery County, Tennessee**
- **Multnomah County, Oregon**
- **Nantucket County, Massachusetts**
- **King County, Washington**
- **Bartholomew County, Indiana**
- **Iredell County, North Carolina**
- **Tulsa County, Oklahoma**
- **Greene County, North Carolina**
- **Marion County, Kentucky**
- **Lauderdale County, Tennessee**
- **Escambia County, Florida**
- **Habersham County, Georgia**
- **Jefferson County, Alabama**
- **Hamilton County, Ohio**
- **Lucas County, Ohio**
- **Madison County, Tennessee**
- **Oklahoma County, Oklahoma**
- **Chatham County, Georgia**
- **Portsmouth city, Virginia**
- **Columbia County, Arkansas**
- **Kemper County, Mississippi**
- **Hampton city, Virginia**
- **Putnam County, Indiana**
- **Clinton County, New York**
- **Dougherty County, Georgia**
- **Dover County, Delaware**
- **Sumter County, Georgia**
- **Coconino County, Arizona**
- **Bannock County, Idaho**
- **Lumpkin County, Georgia**
- **Pickens County, South Carolina**
- **Putnam County, Tennessee**
- **Crawford County, Kansas**
- **Erath County, Texas**
- **Webb County, Texas**
- **Barranquitas Municipio, Puerto Rico**
- **Loíza Municipio, Puerto Rico**
- **Camuy Municipio, Puerto Rico**
- **Florida Municipio, Puerto Rico**
- **Río Grande Municipio, Puerto Rico**
- **Aibonito Municipio, Puerto Rico**
- **Yauco Municipio, Puerto Rico**
- **Barceloneta Municipio, Puerto Rico**
- **Juncos Municipio, Puerto Rico**
- **Prowers County, Colorado**
- **Clay County, Mississippi**
- **Lea County, New Mexico**
- **Ector County, Texas**
- **Nueces County, Texas**
- **Calcasieu Parish, Louisiana**
- **Meade County, South Dakota**
- **Pennington County, South Dakota**
- **Jackson County, South Dakota**
- **Millard County, Utah**
- **Cassia County, Idaho**
- **Beaver County, Utah**
- **DeSoto County, Florida**
- **Beaufort County, South Carolina**
- **Barceloneta Municipio, Puerto Rico**
- **Juncos Municipio, Puerto Rico**
- **Camuy Municipio, Puerto Rico**
- **Florida Municipio, Puerto Rico**
- **Río Grande Municipio, Puerto Rico**
- **Aibonito Municipio, Puerto Rico**
- **Yauco Municipio, Puerto Rico**
The classical k-means approach based on $\delta_1$ gives uninteresting results – the clusters have a single peak. The peak value prevails over other smaller values in the distributions. Using $\delta_3$ as the basic dissimilarity we obtained much more interesting results [Kejžar, N. et al. (2011)].
Patents clusters for $\delta_3$
Patents / clustering of leaders

Symbolic data analysis

V. Batagelj

Clustering and optimization

Leaders method

Agglomerative method

Examples

References

V. Batagelj

Symbolic data analysis


References II


