



EDA, SDA

V. Batagelj

Symbolic data
analysis

Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References

Exploratory data analysis

Symbolic data analysis

Vladimir Batagelj

IMFM Ljubljana, IAM UP Koper, NRU HSE Moscow

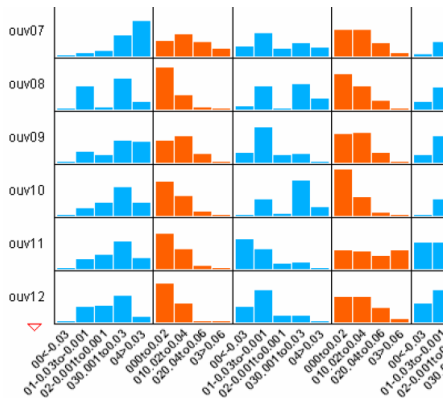
Master's programme

Applied Statistics with Social Network Analysis

International Laboratory for Applied Network Research

NRU HSE, Moscow 2021

- 1 Symbolic data analysis
- 2 Clustering and optimization
- 3 Leaders method
- 4 Agglomerative method
- 5 Examples
- 6 References



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

Current version of slides (November 21, 2021 at 23:24): [slides PDF](#)



Symbolic data

EDA, SDA

V. Batagelj

Symbolic data analysis

Clustering and optimization

Leaders method

Agglomerative method

Examples

References

Data table

	...	variable _{<i>j</i>}	...
...
unit _{<i>i</i>}	...	value _{<i>i,j</i>}	...
...

In classical data analysis value_{*i,j*} is a single element (number or label) measured in standard measurement scales (absolute, ratio, interval, ordinal, nominal).

In *symbolic data table* value_{*i,j*} 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

EDA, SDA

V. Batagelj

Symbolic data analysis

Clustering and optimization

Leaders method

Agglomerative method

Examples

References

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. \(1987\)](#), [Diday, E. \(1995\)](#), [Diday, E. \(2016\)](#)].

Three European projects:

- SODAS - Symbolic Official Data Analysis System (1996-99),
- ISO-3D - Interpretation of Symbolic Objects with 3D Representation (1998-01),
- ASSO – Analysis System of Symbolic Official Data (2001-03).

resulted in a program for symbolic data analysis **SODAS 2**.

The results were published in many papers in conference proceedings and scientific journals, and three books [[Bock, H-H., Diday, E. \(2000\)](#), [Billard, L., Diday, E. \(2006\)](#), [Diday, E., Noirhomme, M. \(2008\)](#)]. Additional two books are to appear soon.



Symbolic data analysis

EDA, SDA

V. Batagelj

Symbolic data analysis

Clustering and optimization

Leaders method

Agglomerative method

Examples

References

The SDA group regularly meets at workshops: Wienerwaldhof (2009), Namur (2011), Beijing (2011), Madrid (2012), Taipei (2014), Orléans (2015), [Ljubljana](#) (2017), [Viana do Castelo](#) (2018).

Three packages for SDA are available in R:

- [RSDA](#),
- [symbolicDA](#),
- [Clamix](#)

[Symbolic data analysis group at LinkedIn](#)



Symbolic data analysis and big data

EDA, SDA

V. Batagelj

Symbolic data analysis

Clustering and optimization

Leaders method

Agglomerative method

Examples

References

big data $\xrightarrow{\text{aggregation}}$ symbolic data table

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 Kežžar) we are developing the clustering algorithms for symbolic objects described by modal valued symbolic data

[Korenjak-Černe, S., Batagelj, V. (1998), Korenjak-Černe, S., Batagelj, V.. (2002), Batagelj, V. et al. (2014), Kežžar, N. et al. (2021)].





Clustering symbolic data

EDA, SDA

V. Batagelj

Symbolic data
analysis

Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References

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)]).
- Ward's *hierarchical clustering method* [Ward, J. H. (1963)].

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

EDA, SDA

V. Batagelj

Symbolic data analysis

Clustering and optimization

Leaders method

Agglomerative method

Examples

References

An SO X is described by a list $X = [\mathbf{x}_i]$ of descriptions of variables V_j . 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

$$\mathbf{f}_{xi} = [f_{xi1}, f_{xi2}, \dots, f_{xik_j}].$$

With

$$\mathbf{x}_i = [p_{xi1}, p_{xi2}, \dots, p_{xik_j}]$$

we denote the corresponding probability distribution.

$$\sum_{j=1}^{k_j} p_{xij} = 1, \quad i = 1, \dots, m$$



Clustering and optimization

EDA, SDA

V. Batagelj

Symbolic data
analysis

Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References

We approach the clustering problem as an optimization problem over the set of *feasible* clusterings Φ_k – partitions of units into k clusters. The *criterion function* has the following form

$$P(\mathbf{C}) = \sum_{C \in \mathbf{C}} p(C). \quad (1)$$

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

We assume a model in which the error of a cluster is a sum of differences of its units from the cluster's *representative* T

$$p(C, T) = \sum_{X \in C} d(X, T). \quad (2)$$

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



Representatives

EDA, SDA

V. Batagelj

Symbolic data
analysis

Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References

The best representative is called a *leader*

$$T_C = \operatorname{argmin}_T p(C, T). \quad (3)$$

Then we define

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

The SO X is described by a list $X = [\mathbf{x}_j]$. Assume that also representatives are described in the same way $T = [\mathbf{t}_j]$,
 $\mathbf{t}_j = [t_{j1}, t_{j2}, \dots, t_{jk_j}]$.



Dissimilarity between SOs

EDA, SDA

V. Batagelj

Symbolic data
analysis

Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References

We introduce a dissimilarity measure between SOs with

$$d(X, T) = \sum_i \alpha_i d(\mathbf{x}_i, \mathbf{t}_i), \quad \alpha_i \geq 0, \quad \sum_i \alpha_i = 1, \quad (5)$$

where

$$d(\mathbf{x}_i, \mathbf{t}_i) = \sum_{j=1}^{k_i} w_{xij} \delta(p_{xij}, t_{ij}), \quad w_{xij} \geq 0. \quad (6)$$

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

EDA, SDA

V. Batagelj

Symbolic data
analysis

Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References

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:

determine an initial clustering

repeat

determine leaders of the clusters in the current clustering;
assign each unit to the nearest new leader – producing a
new clustering

until the leaders stabilize.

Selection of the new leaders

EDA, SDA

V. Batagelj

Symbolic data
analysis

Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References

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

$$T_C = \operatorname{argmin}_T \sum_{X \in C} d(X, T) = \left[\operatorname{argmin}_{\mathbf{t}_i} \sum_{X \in C} d(\mathbf{x}_i, \mathbf{t}_i) \right]_{i=1}^m$$

Therefore $T_C = [\mathbf{t}_i^*]$ and $\mathbf{t}_i^* = \operatorname{argmin}_{\mathbf{t}_i} \sum_{X \in C} d(\mathbf{x}_i, \mathbf{t}_i)$. To simplify the notation we omit the index i .

$$\mathbf{t}^* = \operatorname{argmin}_{\mathbf{t}} \sum_{X \in C} d(\mathbf{x}, \mathbf{t}) = \left[\operatorname{argmin}_{t_j \in \mathbb{R}} \sum_{X \in C} w_{xj} \delta(\rho_{xj}, t_j) \right]_{j=1}^k$$



Leaders

EDA, SDA

V. Batagelj

Symbolic data
analysis

Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References

Again we omit the index j

$$t^* = \operatorname{argmin}_{t \in \mathbb{R}} \sum_{X \in \mathcal{C}} w_X \delta(p_X, t)$$

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

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

or

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



Dissimilarities δ

EDA, SDA

V. Batagelj

Symbolic data analysis

Clustering and optimization

Leaders method

Agglomerative method

Examples

References

	$\delta(x, t)$	t_{ij}^*
1	$(p_x - t)^2$	$\frac{P_{ij}}{A_{ij}}$
2	$(\frac{p_x - t}{t})^2$	$\frac{Q_{ij}}{P_{ij}}$
3	$\frac{(p_x - t)^2}{t}$	$\sqrt{\frac{Q_{ij}}{A_{ij}}}$
4	$(\frac{p_x - t}{p_x})^2$	$\frac{H_{ij}}{F_{ij}}$
5	$\frac{(p_x - t)^2}{p_x}$	$\frac{A_{ij}}{H_{ij}}$
6	$\frac{(p_x - t)^2}{p_x t}$	$\sqrt{\frac{P_{ij}}{H_{ij}}}$

$$A_{ij} = \sum_{X \in C} w_{xij} \quad P_{ij} = \sum_{X \in C} w_{xij} p_{xij} \quad Q_{ij} = \sum_{X \in C} w_{xij} p_{xij}^2$$

$$H_{ij} = \sum_{X \in C} \frac{w_{xij}}{p_{xii}} \quad F_{ij} = \sum_{X \in C} \frac{w_{xij}}{p_{xii}^2}$$

V. Batagelj

EDA, SDA



Leaders

EDA, SDA

V. Batagelj

Symbolic data
analysis

Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References

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 δ_1

EDA, SDA

V. Batagelj

Symbolic data
analysis

Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References

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

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

The leaders' components are *distributions*.

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

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

The leader of a cluster is its *distribution*.



Determining the new clustering

EDA, SDA

V. Batagelj

Symbolic data
analysis

Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References

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

$$P(\mathbf{C}^*) = \sum_{X \in \mathcal{U}} \min_{T \in \mathbf{T}} d(X, T) = \sum_{X \in \mathcal{U}} d(X, T_{c^*(X)}) \quad (8)$$

where

$$c^*(X) = \operatorname{argmin}_k d(X, T_k)$$

We assign each unit X to the closest leader $T_k \in \mathbf{T}$.



Hierarchical agglomerative clustering

EDA, SDA

V. Batagelj

Symbolic data
analysis

Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References

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

each unit forms a cluster: $\mathbf{C}_n = \{\{X\}: X \in \mathcal{U}\}$;

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

$(u, v) = \operatorname{argmin}_{i,j: i \neq j} \{D(C_i, C_j) : C_i, C_j \in \mathbf{C}_{k+1}\}$;

 join the closest pair of clusters $C_{(uv)} = C_u \cup C_v$

$\mathbf{C}_k = (\mathbf{C}_{k+1} \setminus \{C_u, C_v\}) \cup \{C_{(uv)}\}$;

$h(C_{(uv)}) = D(C_u, C_v)$

 determine the dissimilarities $D(C_{(uv)}, C_s), C_s \in \mathbf{C}_k$

endfor

\mathbf{C}_k is a partition of the finite set of units \mathcal{U} into k clusters.

The level $h(C)$ of the cluster $C_{(uv)} = C_u \cup C_v$.





Dissimilarity between clusters

EDA, SDA

V. Batagelj

Symbolic data
analysis

Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References

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

$$\begin{aligned} D(C_U, C_V) &= p(C_U \cup C_V) - p(C_U) - p(C_V) \\ &= \sum_i \alpha_i \sum_j \frac{A_{uij} \cdot A_{vij}}{A_{uij} + A_{vij}} (u_{ij} - v_{ij})^2 \end{aligned} \quad (9)$$

a *generalized Ward's relation*. \mathbf{u}_i and \mathbf{v}_i are components of the leaders of clusters C_U and C_V .



Other dissimilarities

EDA, SDA

V. Batagelj

Symbolic data analysis

Clustering and optimization

Leaders method

Agglomerative method

Examples

References

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).



Scheme of analysis

EDA, SDA

V. Batagelj

Symbolic data analysis

Clustering and optimization

Leaders method

Agglomerative method

Examples

References

raw data



ENCODE



unified data



MAKE SOs



SOs - lists of distributions



leaderSO



clustering and cluster leaders



hclustSO



hierarchy and cluster leaders



ANALYSIS



dendrogram, reports



Population pyramids / World 2001

EDA, SDA

V. Batagelj

Symbolic data analysis

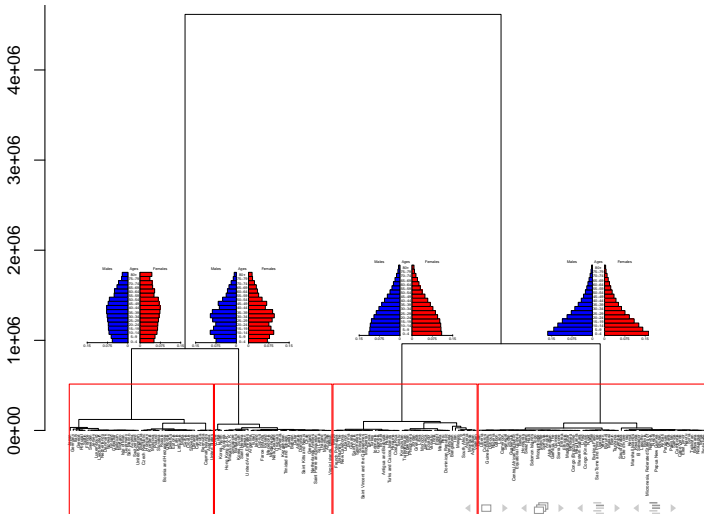
Clustering and optimization

Leaders method

Agglomerative method

Examples

References



V. Batagelj

EDA, SDA

World 2001 / 4 clusters on the map

EDA, SDA

V. Batagelj

Symbolic data analysis

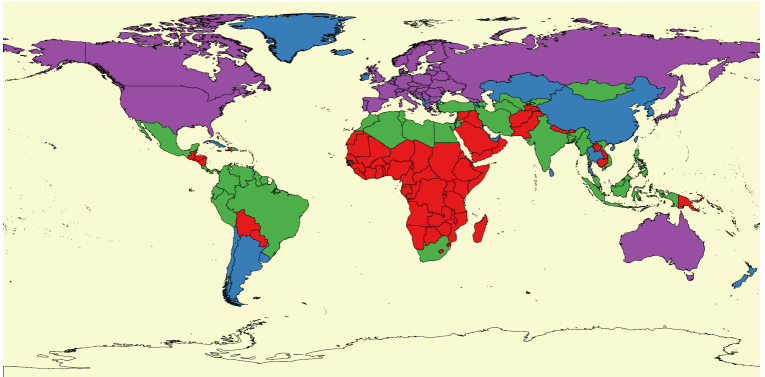
Clustering and optimization

Leaders method

Agglomerative method

Examples

References



[Korenjak-Černe, S. et al. (2015)]



Population pyramids / US counties

EDA, SDA

V. Batagelj

Symbolic data analysis

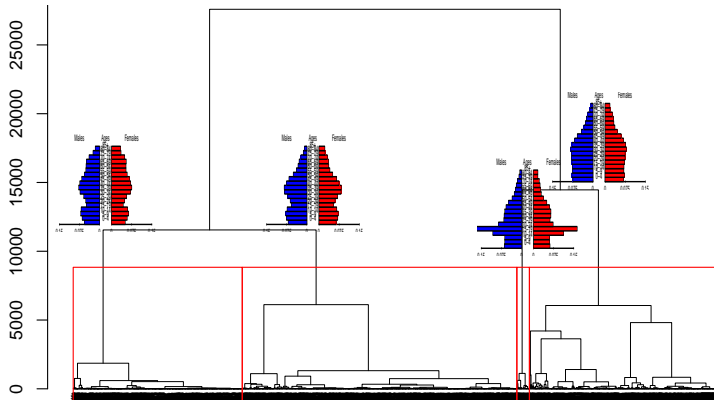
Clustering and optimization

Leaders method

Agglomerative method

Examples

References





US counties map

EDA, SDA

V. Batagelj

Symbolic data analysis

Clustering and optimization

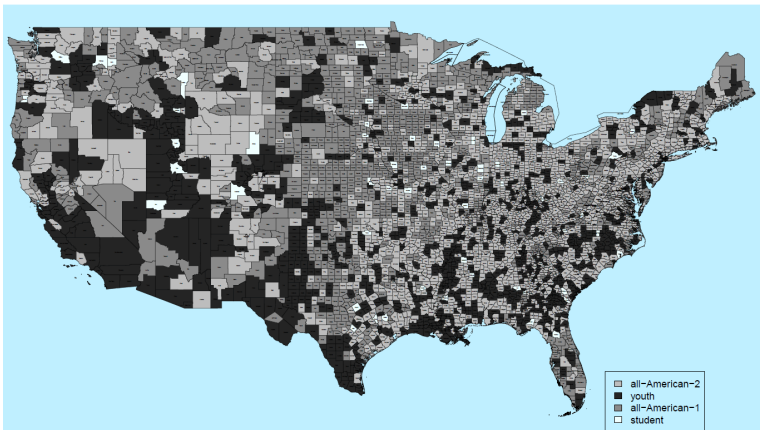
Leaders method

Agglomerative method

Examples

References

US counties PP clustering into 4 clusters





k-means clusters of patents with patents' temporal distributions and cluster leaders

EDA, SDA

V. Batagelj

Symbolic data analysis

Clustering and optimization

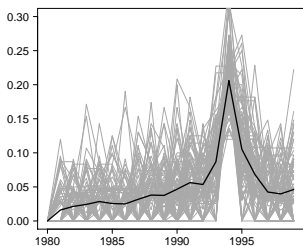
Leaders method

Agglomerative method

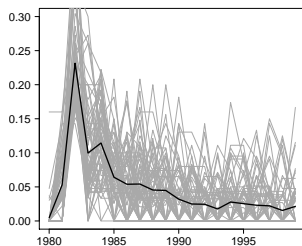
Examples

References

patents: 89 ; Cluster 4



patents: 66 ; Cluster 18



The classical k-means approach based on δ_1 gives uninteresting results – the clusters have a single peak. The peak value prevails over other smaller values in the distributions. Using δ_3 as the basic dissimilarity we obtained much more interesting results [[Kejžar, N. et al. \(2011\)](#)].



Patents clusters for δ_3

EDA, SDA

V. Batagelj

Symbolic data analysis

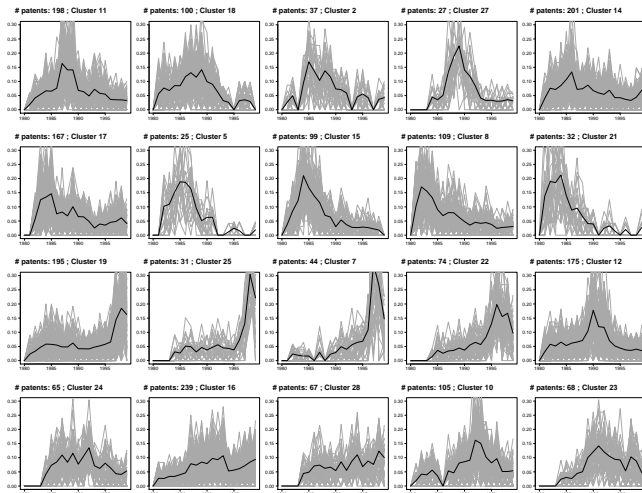
Clustering and optimization

Leaders method

Agglomerative method

Examples

References





Patents / clustering of leaders

EDA, SDA

V. Batagelj

Symbolic data analysis

Clustering and optimization

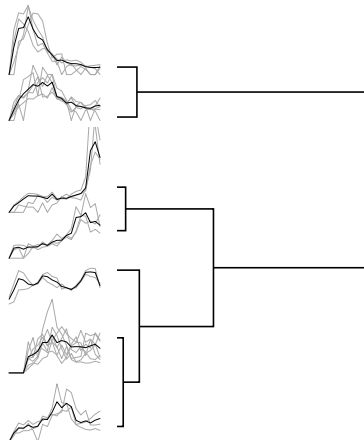
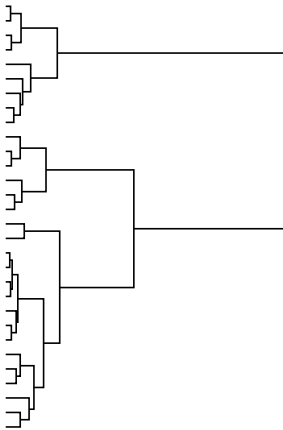
Leaders method

Agglomerative method

Examples

References

C5
C15
C8
C21
C17
C11
C18
C2
C14
C7
C19
C25
C13
C1
C22
C4
C20
C29
C30
C26
C28
C23
C24
C27
C3
C6
C9
C12
C10
C16





References I

EDA, SDA

V. Batagelj

Symbolic data
analysis








Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References

-  Anderberg, M. R. (1973). *Cluster Analysis for Applications*. Academic Press: New York.
-  Batagelj, V. (1988). *Generalized Ward and Related Clustering Problems*. Classification and Related Methods of Data Analysis. H.H. Bock (editor). North-Holland, Amsterdam, 1988. p. 67-74.
-  Batagelj, V., Kejžar, N., Korenjak-Černe, S. (2014) Clustering of modal valued symbolic data. Submitted. [arXiv](#)
-  Billard, L., Diday, E. (2006). *Symbolic data analysis. Conceptual statistics and data mining*. Wiley: New York.
-  Bock, H-H., Diday, E. (eds.) (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Berlin.
-  Diday, E. (1979). *Optimisation en classification automatique*, Tome 1.,2.. INRIA, Rocquencourt (in French).
-  Diday, E. (1987). Introduction à l'approche symbolique en analyse des données. Première Journées 'Symbolique-Numerique'. CEREMADE, Université Paris IX - Dauphine, 21-56.



References II

EDA, SDA

V. Batagelj

Symbolic data
analysis

Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References



Diday, E. (1995). Probabilist, possibilist and belief objects for knowledge analysis. *Annals of Operations Research*. 55, pp. 227–276.



Diday, E. (2016). Thinking by classes in data science: the symbolic data analysis paradigm. *WIREs Comp Stat* 2016, 8: 172-205.



Diday, E. and Noirhomme, M. (eds.) (2008). *Symbolic Data and the SODAS Software*. John Wiley & Sons, Ltd, Chichester.



Hartigan, J. A. (1975). *Clustering algorithms*, Wiley-Interscience: New York.



Japelj Pavešić, B., Korenjak-Černe, S. (2004). "Differences in teaching and learning mathematics in classes over the world : the application of adapted leaders clustering method". In *Proceedings of the IRC - 2004 : IEA International Research Conference*. Nicosia: University of Cyprus, Department of Education, 2004, p. 85–101.



Kejžar, N., Korenjak-Černe, S., Batagelj, V. (2011). "Clustering of discrete distributions: A case of patent citations". *J. classif.*, vol. 28, no. 2, p. 156-183.



Kejžar, N., Korenjak-Černe, S., Batagelj, V. (2021): Clustering of modal-valued symbolic data. *Advances in Data Analysis and Classification*, 15, 513–541.

References III

EDA, SDA

V. Batagelj

Symbolic data
analysis

Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References



Korenjak-Černe, S., Kejžar, N., Batagelj, V. (2015). A weighted clustering of population pyramids for the world's countries, 1996, 2001, 2006. *Population Studies: A Journal of Demography*, 69(1), 105-120.



Korenjak-Černe, S., Batagelj, V., Japelj Pavešić, B. (2011). Clustering large data sets described with discrete distributions and its application on TIMSS data set. *Stat. anal. data min.*, vol. 4, iss. 2, p. 199-215.



Korenjak-Černe, S., Kogovšek, T., Batagelj, V. (2000). Clustering ego-centered networks. In: Blasius, Jörg (ed.). *Social science methodology in the new millennium: proceedings of the Fifth International Conference on Logic and Methodology*. Cologne: TT-Publikaties, 4 pages.



Korenjak-Černe, S., Batagelj, V. (1998). Clustering large datasets of mixed units. In: Rizzi, A., Vichi, M., Bock, H-H. (eds.). *6th Conference of the IFCS, Università "La Sapienza", Rome, 21-24 July, 1998. Advances in data science and classification*. Berlin: Springer, p. 43-48.



Korenjak-Černe, S., Batagelj, V. (2002). Symbolic data analysis approach to clustering large datasets. In: Jajuga, K., Sokołowski, A., Bock, H-H. (eds.). *8th Conference of the IFCS, July 16-19, 2002, Cracow, Poland. Classification, clustering and data analysis*. Berlin: Springer, p. 319-327.



References IV

EDA, SDA

V. Batagelj

Symbolic data
analysis

Clustering and
optimization

Leaders
method

Agglomerative
method

Examples

References



Ward, J. H. (1963). "Hierarchical grouping to optimize an objective function", *Journal of the American Statistical Association*, 58, 236–244.