Symbolic data analysis approach to clustering large datasets

Simona Korenjak-Černe,

Vladimir Batagelj

University of Ljubljana,

Slovenia

IFCS 2002

July 15-19, 2002,

Cracow, Poland

Table of contents

Introduction

Lots of large datasets are available in databases. For the description of the data vector descriptions are usually used. Each its component corresponds to a variable which can be measured in different scales (nominal, ordinal, or numeric).

Most of the well known clustering methods are implemented only for numerical data (e.g., k-means method) or are too complex for clustering large datasets (such as hierarchical methods based on dissimilarity matrices).

For these reasons we propose to use for clustering large datasets a combination of the adapted leaders and hierarchical clustering methods based on special descriptions of units and clusters with the distributions.

The descriptions of units and clusters

Let E be a finite set of units $\mathbf{X}$, which are described by frequency/probability distributions of their descriptors $\{V_1, \cdots, V_m\}$ (Korenjak-Černe and Batagelj (1998)).

The domain of each variable V is partitioned into k_V sub-sets $\{V_i,\ i = 1, \ldots k_V\}$.

For a cluster C we denote

$$
\begin{aligned}
Q(i, C; V) &= \{\mathbf{X} \in C : V(\mathbf{X}) \in V_i\},\ i = 1, \ldots, k_V, \\
q(i, C; V) &= \mathrm{card}(Q(i, C; V)),\ \text{(frequency)} \\
f(i, C; V) &= \frac{q(i, C; V)}{\mathrm{card}(C)},\ \text{(relative frequency)}
\end{aligned}
$$

where $V(\mathbf{X})$ is the value of variable V on unit $\mathbf{X}$, and $\mathrm{card}(C)$ is the number of units in the cluster C.

It holds

$$\sum_{i=1}^{k_V} f(i, C; V) = 1$$

Such a description has the following important properties:

- it requires a **fixed space** per variable;
- it is **compatible with merging** of disjoint clusters – knowing the description of clusters C_1 and C_2, $C_1 \cap C_2 = \emptyset$, we can, without additional information, produce the description of their union

$$f(i, C_1 \cup C_2; V) = \frac{\mathrm{card}(C_1)\ f(i, C_1; V) + \mathrm{card}(C_2)\ f(i, C_2; V)}{\mathrm{card}(C_1 \cup C_2)};$$

- it produces an **uniform description** for all the types of descriptors.

Clustering as an optimization problem

Suppose that our units descriptions consist of m selected variables from $\mathbf{X}$. We define the dissimilarity between two units $\mathbf{X}_1$ and $\mathbf{X}_2$ as the weighted sum of the dissimilarity between them on each variable V

$$d(\mathbf{X}_1, \mathbf{X}_2) = \sum_{j=1}^{m} \alpha_j \, d(\mathbf{X}_1, \mathbf{X}_2; V_j), \qquad \sum_{j=1}^{m} \alpha_j = 1$$

where

$$d_{abs}(\mathbf{X}_1, \mathbf{X}_2; V) = \frac{1}{2} \sum_{i=1}^{k(V)} |f(i, \mathbf{X}_1; V) - f(i, \mathbf{X}_2; V)|$$

or

$$d_{sqr}(\mathbf{X}_1, \mathbf{X}_2; V) = \frac{1}{2} \sum_{i=1}^{k(V)} (f(i, \mathbf{X}_1; V) - f(i, \mathbf{X}_2; V))^2.$$

Here, $\alpha_j \geq 0$ $(j = 1, \ldots, m)$ denote weights, which could be equal for all variables or different if we have same information about their importance. Because the clusters are represented in the same way we also use so defined dissimilarity for defining the dissimilarity between two clusters.

On the basis of this definition we define clustering problem as an optimization problem:

Find clustering $\mathbf{C}$ for which

$$P(\mathbf{C}^*) = \min_{\mathbf{C} \in \Phi} P(\mathbf{C}).$$

For the criterion function as a measure of goodness of clustering we use its most common definition

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

and

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

where L_C represents the leader (the representative element) of the cluster C.

The adapted leaders method

For clustering very large datasets where variables of units are measured in different scales we developed an adapted version of the leaders method. The proposed method is a variant of **the dynamic clustering method** and can be shortly described with the following procedure:

determine an initial clustering
repeat
    determine leaders of the clusters;
    assign each unit to the nearest new leader
        – producing a new clustering
until the leaders do not change more.

In the elaboration of the proposed approach we assume that the descriptions of the leaders have the same form as the descriptions of the units and clusters:

$$\begin{aligned} L &= [L(V_1), \ldots, L(V_m)], \\ L(V) &= [s(1, L; V), \ldots, s(k_V, L; V)], \end{aligned}$$

where $\sum_{j=1}^{k_V} s(j, L; V) = 1$

It can be proved that for the first criterion function P_{abs} with the dissimilarity d_{abs}, the optimal leaders are determined with **maximal frequencies**

$$s(i, L; V) = \begin{cases} \frac{1}{t} & \text{if } j \in M \\ 0 & \text{otherwise} \end{cases}$$

where $M = \{j : q(j, C; V) = \max_i q(i, C; V)\}$ and $t = \text{card}(M)$.

The precondition for this result is that all units should be represented with a single value for each variable (this is usually the case).

For the second criterion function P_{sqr} with dissimilarity d_{sqr} the optimal leaders are uniquely determined with the **averages of the relative frequencies**

$$s(i, L; V) = \frac{1}{\text{card}(C)} \sum_{X \in C} f(i, X; V).$$

This is an extended version of the well-known k-means method, which is appropriate only for numerical variables (Hartigan (1975)).

The main advantages of the second method are:

- the input **unit can be represented also with distributions**, and not only with a single value for each variable,
- the optimal **leaders are uniquely determined**,
- if all units are represented with a single value for each variable, the optimal leader is determined with the relative frequencies of the cluster

$$s(i, L; V) = f(i, C; V).$$

Hierarchical clustering method on the leaders

To produce a hierarchical clustering on the leaders the standard **agglomerative hierarchical clustering method** is used:

each unit is a cluster: $\mathbf{C}_1 = \{\{\mathbf{X}\}: \mathbf{X} \in E\}$;
they are at level 0: $h(\{\mathbf{X}\}) = 0,\ \mathbf{X} \in E$;
for $k := 1$ **to** $n - 1$ **do**
determine the closest pair of clusters
$(p, q) = \text{argmin}_{i,j:i \neq j}\{D(C_i, C_j): C_i, C_j \in \mathbf{C}_k\}$;
join them
$\mathbf{C}_{k+1} = \mathbf{C}_k \setminus \{C_p, C_q\} \cup \{C_p \cup C_q\}$;
$h(C_p \cup C_q) = D(C_p, C_q)$
endfor

Graphical representation of the hierarchical clustering method is called a **dendrogram**. The level $h(C)$ of cluster $C = C_p \cup C_q$ is determined by the dissimilarity between the joined clusters C_p and C_q

$$h(C_p \cup C_q) = D(C_p, C_q).$$

$h(C) = 0$ for C is a leader.

For the first criterion function the following formula for the calculation of the dissimilarity between clusters $D(C_p, C_q)$ is used

$$\begin{aligned} D(C_p, C_q) &= P(\mathbf{C}_{k+1}) - P(\mathbf{C}_k) = \\ &= p(C_p \cup C_q) - p(C_p) - p(C_q) \end{aligned}$$

where

$$p(C) = \sum_{\mathbf{X} \in C} d(\mathbf{X}, L_C).$$

For the second criterion function, $D(C_p, C_q)$ is determined from the well known Ward's relation

$$D(C_p, C_q) = \frac{\mathrm{card}(C_p) \cdot \mathrm{card}(C_q)}{\mathrm{card}(C_p) + \mathrm{card}(C_q)} \, d(L_p, L_q)$$

.

An Example - the interpretation of the results

Source:

USDA Nutrient Database for Standard Reference, Release 14. U.S. Department of Agriculture, Agricultural Research Service. 2001: Nutrient Data Laboratory Home Page, `http://www.nal.usda.gov/fnic/foodcomp`.

The dataset contains data on 6039 foods – units. We considered in this study 31 nutrients – numerical variables describing each food. This dataset was selected because the results can be interpreted in easy-to-understand way.

6.1 Partition of domains of variables

The domain of each variable is divided into 10 sub-sets: one with the value, that indicates missing value, one with the value zero, and one special sub-set with outlying (extremely large) values. The rest of the values are divided into 7 sub-sets with equal number of values (Dougherty et al. (1995)).

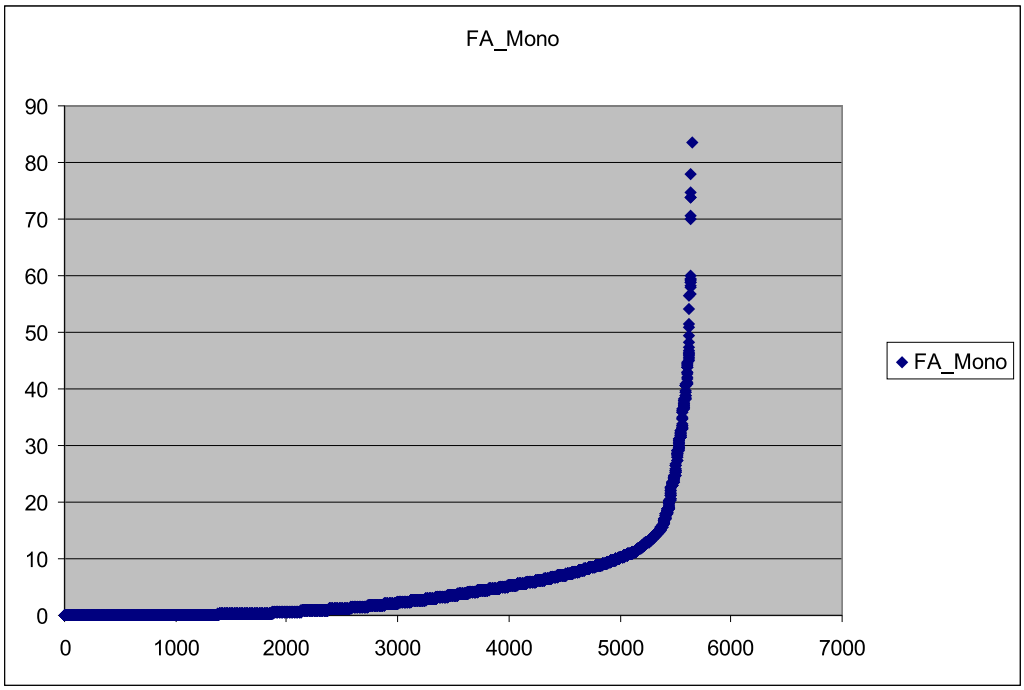

Figure 1: The graph of the distribution of the variable *fa-mono*.

The partition of it's domain:

$$\text{var} = fa - mono$$

MAP

$$\begin{aligned} -1\ (missing) \qquad & 5 = [1.25, 3] \\ 1 = \{0\} \qquad & 6 = (3, 5.6) \\ 2 = (0, 0.035] \qquad & 7 = [5.6, 9.5) \\ 3 = (0.035, 0.3] \qquad & 8 = [9.5, 65) \\ 4 = (0.3, 1.25) \qquad & 9 = [65, 85]\ (outliers) \end{aligned}$$

6.2 Transformed data

Each unit is represented with the vector of indices of sub-sets in which lie its real values. For example: the food BUTTER,WITH SALT with the ID = 1001 has for the first four variables and their indices the following values:

ID	water	food energy	protein	total lipid (fat)
1001	15.87	717	0.85	81.11
1001	3	8	2	8

because for the *water* (g/100g) the third sub-set is $3 = (5.65, 29.5]$, the eighth sub-set for *energy* (kcal/100g) is $8 = (386, 800)$, the second sub-set for *protein* (g/100g) is $2 = (0, 1.5]$, and the eighth sub-set for *fat* (g/100g) is $8 = [23.5, 85)$.

6.3 Clustering results

In the leaders program the initial clustering with 30 clusters was randomly selected. For the selected dissimilarity d_{sqr} the 30 leaders stabilized after 29 iterations.

The hierarchy we got has three main branches: meats, (mainly) vegetables, and (mainly) cereals.
For each node of the dendrogram, the distribution for each variable is also determined.

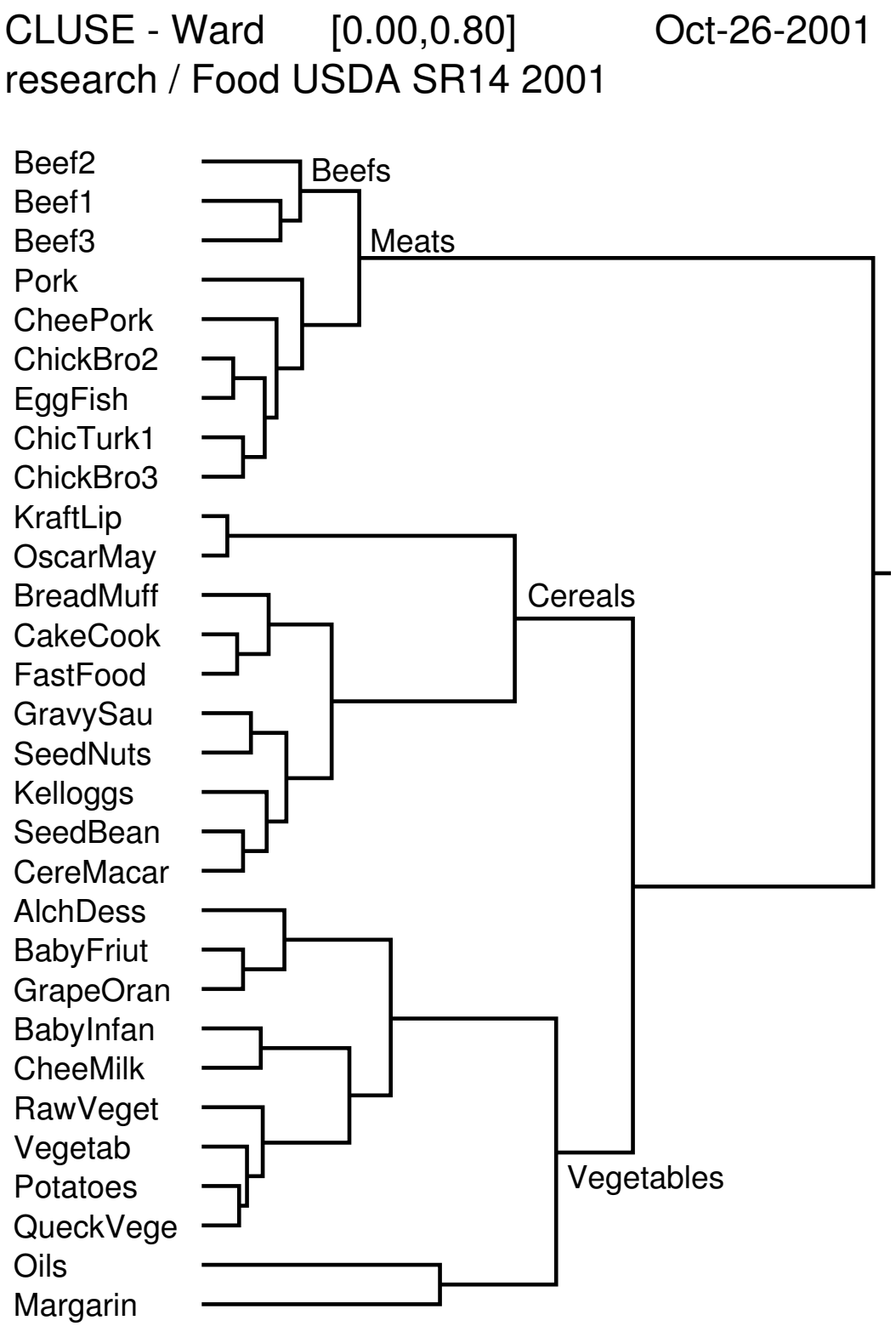

Figure 2: The dendrogram on 30 leaders of food's clusters.

Table 1: $q(Beefs, V)$

V	{0}	2	3	4	5	6	7	8	out	mis
$water$	0	0	1	167	**264**	130	29	0	0	0
$energ-kc$	0	0	0	54	**231**	183	103	20	0	0
$protein$	0	0	0	0	8	245	**338**	0	0	0
$tot-lipi$	0	0	0	10	101	**178**	169	133	0	0
$carbohyd$	**587**	3	1	0	0	0	0	0	0	0
$fiber-td$	**588**	0	0	0	0	0	0	0	0	3
ash	0	0	195	**278**	101	13	4	0	0	0
$calcium$	0	**294**	241	52	2	0	0	2	0	0
$phosphor$	0	0	0	0	71	227	**273**	20	0	0
$iron$	0	0	0	1	61	248	**277**	4	0	0
$sodium$	0	0	26	**413**	148	0	0	4	0	0
$potassiu$	0	0	0	0	75	225	**279**	12	0	0
$magnesiu$	0	0	11	166	197	**208**	9	0	0	0
$zinc$	0	0	0	0	0	160	**431**	0	0	0
$copper$	0	0	76	0	**436**	79	0	0	0	0
$manganes$	0	**346**	242	3	0	0	0	0	0	0
$selenium$	0	0	0	4	168	**378**	41	0	0	0
$vit-A$	**588**	0	0	0	0	0	0	0	0	3
$vit-E$	0	1	**218**	65	0	0	0	0	0	307
$thiamin$	0	0	3	122	**315**	150	0	1	0	0
$ribolfla$	0	0	0	41	**233**	202	115	0	0	0
$niacin$	0	0	0	5	**327**	240	19	0	0	0
$panto-ac$	0	0	11	**375**	201	2	0	0	0	2
$vit-B6$	0	0	0	1	2	188	**311**	89	0	0
$folate$	0	1	**335**	228	23	1	0	0	0	3
$vit-B12$	0	0	0	0	1	163	**300**	127	0	0
$vit-C$	**587**	0	0	0	0	0	0	0	0	4
$fa-sat$	0	0	0	2	65	147	183	**194**	0	0
$fa-mono$	0	0	0	2	81	**184**	170	154	0	0
$fa-poly$	0	2	147	**237**	187	18	0	0	0	0
$cholestr$	0	0	3	150	169	**205**	64	0	0	0

For example, the cluster *Beefs* = *Beef1* ∪ *Beef3* ∪ *Beef2* has the description given in Table 1. It consists of 591 units. From this table the following characteristics of the cluster *Beefs* can be seen:

V	modus sub-set	% of units	extended sub-set	% of units
$fiber-td$	$\{0\}$	99.49	$\{0\} \cup \{missing\}$	100.00
$vit-A$	$\{0\}$	99.49	$\{0\} \cup \{missing\}$	100.00
$vit-C$	$\{0\}$	99.32	$\{0\} \cup \{missing\}$	100.00
$carbohyd$	$\{0\}$	99.32	$[0, 6.85]$	100.00
$zinc$	$[4.12, 20)$	72.93	$[2.23, 20)$	100.00
$sodium$	$[50, 66)$	69.88	$[10, 121)$	99.32

Detailed results and programs are available at `http://www.educa.fmf.uni-lj.si/datana/`.

Conclusion

In our opinion the main advantages of our approach are:

- appropriate for **large** datasets with **mixed** units;
- the units can be **described with distributions** of varible's values, not only with a single value;
- variables can be **weighted** by their importance.
- for each cluster **all distributions** of values for all variables **are stored** during clustering process which provide us with rather simple interpretations of the clustering results.

At URL

`http://www.educa.fmf.uni-lj.si/datana/`

you can find the CLAMIX programs and some of the results.

`e-mail: Simona.Cerne@uni-lj.si`
`e-mail: Vladimir.Batagelj@uni-lj.si`