Symbolic data analysis approach to clustering large datasets

Simona Korenjak-Černe[1] and Vladimir Batagelj[2]

[1] University of Ljubljana, Faculty of Economics,
Kardeljeva ploščad 17, 1 101 Ljubljana, Slovenia,
and IMFM Ljubljana, Department of TCS,
Jadranska ulica 19, 1 000 Ljubljana, Slovenia
(e-mail: simona.cerne@uni-lj.si)
[2] University of Ljubljana, FMF, Department of Mathematics,
and IMFM Ljubljana, Department of TCS,
Jadranska ulica 19, 1 000 Ljubljana, Slovenia
(e-mail: vladimir.batagelj@uni-lj.si)

Abstract. The paper builds on the representation of units/clusters with a special type of symbolic objects that consist of distributions of variables. Two compatible clustering methods are developed: the *leaders method*, that reduces a large dataset to a smaller set of symbolic objects (clusters) on which a *hierarchical clustering method* is applied to reveal its internal structure. The proposed approach is illustrated on *USDA Nutrient Database*.

1 Introduction

Nowadays lots of large datasets are available in databases. One of possible ways how to extract information from these datasets is to find homogeneous clusters of similar units. 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. This description is based on a special kind of symbolic objects (Bock and Diday (2000)), formed by the distributions of partitioned variables over a cluster – histograms.

2 The descriptions of units and clusters

Let E be a finite set of units 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

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

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

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

The description of the cluster C by the variable V is the vector of the frequencies of V_i $(i = 1, \cdots, k_V)$. A unit is considered as a special cluster with only one element and can be in our approach represented either with a single value or by the distributions of the partitioned variables.

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

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

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

3 Dissimilarity

In the following we shall use two dissimilarities, both defined as a weighted sum of the dissimilarities on each variable:

$$d(C_1, C_2) = \sum_{j=1}^{m} \alpha_j\ d(C_1, C_2; V_j), \qquad \sum_{j=1}^{m} \alpha_j = 1 \tag{1}$$

where

$$d_{abs}(C_1, C_2; V_j) = \frac{1}{2} \sum_{i=1}^{k_j} |f(i, C_1; V_j) - f(i, C_2; V_j)| \tag{2}$$

or

$$d_{sqr}(C_1, C_2; V_j) = \frac{1}{2} \sum_{i=1}^{k_j} (f(i, C_1; V_j) - f(i, C_2; V_j))^2, \tag{3}$$

$k_j = k_{V_j}$. 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 the importance of the variables.

For the dissimilarity d_{abs} the triangle inequality also holds. Therefore it is also a semidistance.

4 The adapted leaders method

For clustering large datasets the clustering procedures based on dissimilarity matrix are too time consuming. A more appropriate approach is the adapted leaders method – a variant of the dynamic clustering method (Diday (1979), Korenjak-Černe and Batagelj (1998), Verde et al. (2000)). This method can be shortly described with the following 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 do not change more.

The leaders method is solving the following optimization problem:

Find a clustering $\mathbf{C}^*$ in a set of feasible clusterings Φ for which

$$P(\mathbf{C}^*) = \min_{\mathbf{C} \in \Phi} P(\mathbf{C}) \tag{4}$$

with the criterion function

$$P(\mathbf{C}) = \sum_{C \in \mathbf{C}} p(C) \qquad \text{and} \qquad p(C) = \sum_{X \in C} d(X, L_C), \tag{5}$$

where L_C represents the leader (a representative element) of the cluster C. In our case the set of feasible clusterings Φ is a set of partitions of the set E. The number of the clusters could be fixed a priori or could be determined with the selection of the maximal allowed dissimilarity between the unit and the nearest leader.

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:

$$L = [L(V_1), \ldots, L(V_m)],$$
$$L(V) = [s(1, L; V), \ldots, s(k_V, L; V)],$$

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

It can be proved that for the first criterion function P_{abs}, where in the definition (5) the dissimilarity d_{abs} is used, 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 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.

5 Building a hierarchy

To produce a hierarchical clustering on the clusters represented with their leaders the standard agglomerative hierarchical clustering method is used:

each unit is a cluster: $\mathbf{C}_1 = \{\{X\}: X \in E\}$;
they are at level 0: $h(\{X\}) = 0,\ X \in E$;
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

The level $h(C)$ of the cluster $C = C_p \cup C_q$ is determined by the dissimilarity between the joint clusters C_p and C_q by $h(C_p \cup C_q) = D(C_p, C_q)$. The units X are the clusters from the initial clustering, represented with their leaders. $h(C) = 0$ for C from the initial clustering.

The dissimilarity between clusters $D(C_p, C_q)$ measures the change of the value of the criterion function produced by the merging of the clusters C_p and C_q

$$D(C_p, C_q) = p(C_p \cup C_q) - p(C_p) - p(C_q) \tag{6}$$

For the second criterion function P_{sqr} the dissimilarity $D(C_p, C_q)$ can be determined using the analogue of the Ward's relation (Batagelj (1988)):

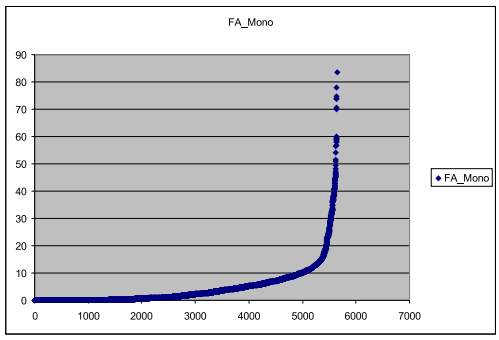
$$D(C_p, C_q) = \frac{\text{card}(C_p) \cdot \text{card}(C_q)}{\text{card}(C_p) + \text{card}(C_q)}\, d(L_p, L_q).$$

6 Example

The proposed approach was successfully applied to some large datasets (for example, the dataset on the topic *Family and Changing Gender Roles I* and *II* with 45 785 units and 33 selected variables from ISSP datasets). We are presenting here the results on the nutrient database from U.S. Department of Agriculture. 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)).




```
var=fa_mono
MAP
-1 (missing)
1={0}
2=(0,0.035]
3=(0.035,0.3]
4=(0.3,1.25)
5=[1.25,3]
6=(3,5.6)
7=[5.6,9.5)
8=[9.5,65)
9=[65,85] (outliers)
```

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

For example, the variable *fa-mono* (total monounsaturated fatty acids) has 395 missing values, 128 units have value 0 and 7 units have extremely large values. The distribution of the values for this variable is presented in the Figure 1 and next to it is our partition of it's domain.

6.2 Transformed data

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

ID	water	food energy	protein	total lipid (fat)	carbohydrate
1001	15.87	717	0.85	81.11	0.06
1001	3	8	2	8	2

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 $2 = (0, 1.5]$, the eighth sub-set for *fat* (g/100g) is $8 = [23.5, 85)$ and the second sub-set for *carbohydrate* (g/100g) is $2 = (0, 3.5]$.

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. On these leaders the hierarchy based on the same dissimilarity was built. The dendrogram displayed in Figure 2 was obtained. 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. For example, the cluster *Beefs* = *Beef1* $\cup$ *Beef3* $\cup$ *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	extanded 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

For each variable the complete distribution can be observed. For example, from the Table 1 we can see that for the variable *fa-mono* 184 (31.13 %) units from this cluster have values in the 6th sub-set $(3, 5.6)$. But if we extend the interval to $(3, 65)$ (union of three sub-sets) 85.96 % of all units from the cluster *Beefs* are included in it. Detailed results and programs are available at `http://www.educa.fmf.uni-lj.si/datana/` .

Acknowledgment

This work was supported by the Ministry of Science and Technology of Slovenia, Project J1-8532.

References

BATAGELJ, V.: *Generalized Ward and related clustering problems.* (H.H. Bock, ed.: Classification and related methods of data analysis), North-Holland, Amsterdam, 1988, 67-74.

BOCK, H.-H. (2000): Symbolic Data. In: H.-H. Bock and E. Diday (Eds.): *Analysis of Symbolic Data.* Exploratory methods for extracting statistical information from complex data. Springer, Heidelberg.

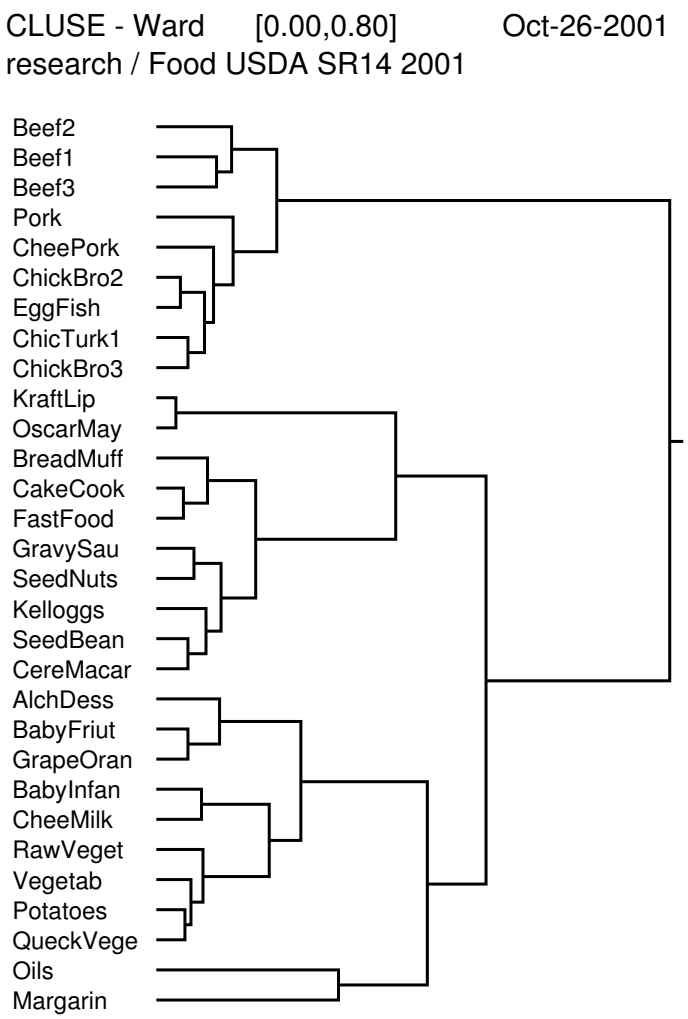


Fig. 2. The dendrogram on 30 leaders of food's clusters.

BOCK, H.-H. and DIDAY, E. (2000): Symbolic Objects. In: H.-H. Bock and E. Diday (Eds.): *Analysis of Symbolic Data*. Exploratory methods for extracting statistical information from complex data. Springer, Heidelberg.

DIDAY, E. (1979): *Optimisation en classification automatique*, Tome 1.,2.. INRIA, Rocquencourt (in French).

DOUGHERTY, J., KOHAVI, R., and SAHAMI, M. (1995): *Supervised and unsupervised discretization of continuous features*. Proceedings of the Twelfth International Conference on Machine Learning (pp. 194–202). Tahoe City, CA: Morgan Kaufmann. `http://citeseer.nj.nec.com/dougherty95supervised.html`

HARTIGAN, J.A. (1975): *Clustering Algorithms*. Wiley, New York.

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 − *B*6	0	0	0	1	2	188	**311**	89	0	0
folate	0	1	**335**	228	23	1	0	0	0	3
vit − *B*12	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

KORENJAK-ČERNE, S. and BATAGELJ, V. (1998): Clustering large datasets of mixed units. In: Rizzi, A., Vichi, M., Bock, H.-H. (Eds.): *Advances in Data Science and Classification.* Springer.

VERDE, R., DE CARVALHO, F.A.T. and LECHEVALLIER, Y. (2000): A Dynamic Clustering Algorithm for Multi-nominal Data. In: Kiers, H.A.L., Rasson, J.-P., Groenen, P.J.F., Schader, M. (Eds.): *Data Analysis, Classification, and Related Methods.* Springer.

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`.