Preprint series, Vol. 40 (2002), 800

SYMBOLIC DATA ANALYSIS
APPROACH TO
CLUSTERING LARGE DATASETS

Simona Korenjak-Černe, Vladimir Batagelj

ISSN 1318-4865

Version 1: October 27, 2001
Version 2: January 11, 2002


Submitted to the IFCS-2002, Cracow, Poland, July 16-19 2002

Supported by the Ministry of Education, Science and Sport of Slovenia, Project J1-8532.

Ljubljana, January 11, 2002
Symbolic data analysis approach to clustering large datasets

Simona Korenjak-Černe
University of Ljubljana, Faculty of Economics,
Kardeljeva ploščad 17, 1101 Ljubljana, Slovenia,
and IMFM Ljubljana, Department of TCS,
Jadranska ulica 19, 1000 Ljubljana, Slovenia
e-mail: simona.cerne@uni-lj.si

Vladimir Batagelj
University of Ljubljana, FMF, Department of Mathematics,
and IMFM Ljubljana, Department of TCS,
Jadranska ulica 19, 1000 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.

Key words: symbolic object, clustering, large data set, distribution.


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 \( \{V_1, \ldots, V_m\} \) (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

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

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 \( V_i \) \( (i = 1, \ldots, k_V) \). 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:

\[ d(C_1, C_2) = \sum_{j=1}^{m} \alpha_j d(C_1, C_2; V_j), \quad \sum_{j=1}^{m} \alpha_j = 1 \] (1)

where

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

or

\[ d_{sqr}(C_1, C_2; V_j) = \frac{1}{2} \sum_{i=1}^{k_j} (f(i, C_1; V_j) - f(i, C_2; V_j))^2, \] (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 \( C^* \) in a set of feasible clusterings \( \Phi \) for which

\[ P(C^*) = \min_{C \in \Phi} P(C) \] (4)
with the criterion function

\[ P(C) = \sum_{C \in \mathcal{C}} p(C) \quad \text{and} \quad p(C) = \sum_{X \in C} d(X, L_C), \quad (5) \]

where \( L_C \) represents the leader (a representative element) of the cluster \( C \). In our case the set of feasible clusterings \( \Phi \) 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:

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

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

It can be proved that for the first criterion function \( P_{obs} \), where in the definition (5) the dissimilarity \( d_{obs} \) 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: $C_1 = \{ \{X \}: X \in E \}$;
- they are at level 0: $h(\{X\}) = 0$, $X \in E$;
- for $k := 1$ to $n - 1$ do
  - determine the closest pair of clusters
  - $(p, q) = \arg\min_{i, j; i \neq j} \{D(C_i, C_j): C_i, C_j \in C_k \}$;
  - join them
  - $C_{k+1} = (C_k \setminus \{C_p, C_q\}) \cup \{C_p \cup C_q\}$;
  - $h(C_p \cup C_q) = D(C_p, C_q)$
- 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 $C_q$

$$D(C_p, C_q) = p(C_p \cup C_q) - p(C_p) - p(C_q)$$  \hspace{1cm} (6)

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

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

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)).

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 the following is our partition of it’s domain:

```
var=fa_mono
MAP
-1 (missing)  5=[1.25,3]
  1={0}       6=(3.5,6)
  2=(0,0.035]  7=[5.6,9.5)
  3=(0.035,0.3] 8=[9.5,65)
  4=(0.3,1.25] 9=[65,85] (outliers)
```
6.2 Transformed data

Each unit is represented with the vector of indices of sub-sets in which it has 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:

<table>
<thead>
<tr>
<th>ID</th>
<th>water</th>
<th>food energy</th>
<th>protein</th>
<th>total lipid (fat)</th>
<th>carbohydrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>3</td>
<td>8</td>
<td>2</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>

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_{sq} 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 Beef = Beef1 ∪ Beef3 ∪ Beef2 has the description given in Table 1. It consists of 591 units. From this table the following characteristics of the cluster Beef can be seen:

<table>
<thead>
<tr>
<th>V</th>
<th>modus sub-set</th>
<th>% of units</th>
<th>extended sub-set</th>
<th>% of units</th>
</tr>
</thead>
<tbody>
<tr>
<td>fiber - td</td>
<td>{0}</td>
<td>99.49</td>
<td>{0} ∪ {missing}</td>
<td>100.00</td>
</tr>
<tr>
<td>vit - A</td>
<td>{0}</td>
<td>99.49</td>
<td>{0} ∪ {missing}</td>
<td>100.00</td>
</tr>
<tr>
<td>vit - C</td>
<td>{0}</td>
<td>99.32</td>
<td>{0} ∪ {missing}</td>
<td>100.00</td>
</tr>
<tr>
<td>carbohyd</td>
<td>{0}</td>
<td>99.32</td>
<td>[0, 6.85]</td>
<td>100.00</td>
</tr>
<tr>
<td>sodium</td>
<td>[50, 66]</td>
<td>69.88</td>
<td>[10, 121]</td>
<td>99.32</td>
</tr>
</tbody>
</table>

For each variable the complete distribution can be observed. For example, from the Table 1 we can see that for the variable fu-monol 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 Beef are included in it. Detailed results and programs are available at

http://www.educa.fmf.uni-lj.si/datana/.
Figure 2: The dendrogram on 30 leaders of food’s clusters.
6.3 Clustering results

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>out</th>
<th>mis</th>
</tr>
</thead>
<tbody>
<tr>
<td>water</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>167</td>
<td>264</td>
<td>130</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>energy kc</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>54</td>
<td>231</td>
<td>183</td>
<td>103</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>protein</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>245</td>
<td>338</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>total lipi</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>101</td>
<td>178</td>
<td>169</td>
<td>133</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>carbohydrates</td>
<td>587</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>fiber td</td>
<td>588</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>ash</td>
<td>0</td>
<td>0</td>
<td>195</td>
<td>278</td>
<td>101</td>
<td>13</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>calcium</td>
<td>0</td>
<td>294</td>
<td>241</td>
<td>52</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>phos</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>71</td>
<td>227</td>
<td>273</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>iron</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>61</td>
<td>248</td>
<td>277</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sodium</td>
<td>0</td>
<td>0</td>
<td>26</td>
<td>413</td>
<td>148</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>potassium</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>225</td>
<td>279</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>magnesium</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>166</td>
<td>197</td>
<td>208</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>zinc</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>160</td>
<td>431</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>copper</td>
<td>0</td>
<td>0</td>
<td>76</td>
<td>0</td>
<td>436</td>
<td>79</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>manganese</td>
<td>0</td>
<td>346</td>
<td>242</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>selenium</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>168</td>
<td>378</td>
<td>41</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>vit A</td>
<td>0</td>
<td>588</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>vit E</td>
<td>0</td>
<td>1</td>
<td>218</td>
<td>65</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>thiamin</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>122</td>
<td>315</td>
<td>150</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ribofl avia</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>41</td>
<td>233</td>
<td>202</td>
<td>115</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>niacin</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>327</td>
<td>240</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>panto ac</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>375</td>
<td>201</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>vit B6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>188</td>
<td>311</td>
<td>89</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>folate</td>
<td>0</td>
<td>1</td>
<td>335</td>
<td>228</td>
<td>23</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>vit B12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>163</td>
<td>300</td>
<td>127</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>nnnn</td>
<td>0</td>
<td>587</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>fa sat</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>65</td>
<td>147</td>
<td>183</td>
<td>194</td>
<td>0</td>
</tr>
<tr>
<td>fa mono</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>81</td>
<td>184</td>
<td>170</td>
<td>154</td>
<td>0</td>
</tr>
<tr>
<td>fa poly</td>
<td>0</td>
<td>2</td>
<td>147</td>
<td>237</td>
<td>187</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>choles</td>
<td>0</td>
<td>3</td>
<td>150</td>
<td>169</td>
<td>205</td>
<td>64</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Acknowledgment

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

References


http://citeseer.nj.nec.com/dougherty95supervised.html


