A divisive approach for clustering symbolic data

M. Chavent

Mathématiques Appliquées de Bordeaux (MAB),
Universités Bordeaux1 et 2
Plan:

1. Divisive clustering method
   - descendant hierarchical algorithm
   - classical or symbolic data

2. Application for clustering a set of categories
   - example of a set of species contaminated with mercury
   - comparison of numerical and symbolic approach for clustering the species
Introduction to divisive clustering

Hierarchical clustering

1. Starts with the whole dataset
2. At each step:
   ✓ Choice of the cluster to be split
   ✓ Split of this cluster
3. Stop:
   ✓ After K iterations or
   ✓ All clusters are singletons or have similar objects
Introduction to divisive clustering

How to split a cluster

✓ Complete enumeration (Edward and Cavalli-Sforza, 1965)
✓ Heuristics (Mac-Naughton-Smith, 1964, Chidananda Gowda et al. 1978)
✓ Criterion optimisation like split or diameter (Guénoche 1991, Wang et al. 1996)
✓ **Monothetic approach**
  ✓ Clustering : Williams & Lambert 1959
  ✓ Classification : Breiman et al. 1984 (CART) or Quinlan 1986 (ID3)
The proposed divisive clustering method

Takes classical or symbolic data as input

\[
X = \left( x_{ij} \right)
\]

\[
x_{ij} = \begin{cases} 
\text{single value} & \text{if } x_{ij} \text{ is a single value} \\
\text{interval of values} & \text{if } x_{ij} \text{ is an interval} \\
\text{set of categories} & \text{if } x_{ij} \text{ is a set of categories} \\
\text{probability distribution} & \text{if } x_{ij} \text{ is a probability distribution}
\end{cases}
\]

- Definition of a dissimilarity measure
- Extension of the within-cluster inertia criterion
The proposed dissimilarity measure

For a numerical symbolic (interval) data matrix :

\[ d^2(x_i, x_{i'}) = \sum_{j=1}^{p} d_H^2(x^j_i, x^j_{i'}) \]

- \( d_H \) is the **Hausdorff distance** between two intervals
- If all intervals are reduced to single points, it corresponds to the Euclidean distance

The normalized version of this distance :

\[ d^2(x_i, x_{i'}) = \sum_{j=1}^{p} \frac{1}{v^j} d_H^2(x^j_i, x^j_{i'}) \quad v^j = \frac{1}{n^2} \sum_{i=1}^{n-1} \sum_{i' > i} d_H^2(x^j_i, x^j_{i'}) \]
The proposed dissimilarity measure

From a symbolic categorical data matrix

\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{height} & \text{weight} \\
\hline
\text{small} & \text{large} & \text{very light} & \text{light} & \text{heavy} \\
\hline
1 & 0.3 & 0.7 & 0.5 & 0.5 & 0 \\
2 & 0.8 & 0.2 & 1 & 0 & 0 \\
3 & 0 & 1 & 0 & 0 & 1 \\
\hline
\end{array}
\]

\[
\begin{array}{cccc}
1 & \ldots & j & \ldots & m \\
\hline
1 & & & & \\
\cdots & & & & \\
i & x_{ij} & & \\
\cdots & & & & \\
n & & & & \\
\hline
\end{array}
\]

\[
d^2(x_i, x_{i'}) = \sum_{j=1}^{p} d_{\Phi2}^2(x_i^j, x_{i'}^j)
\]

\(\Rightarrow d_{\Phi2} \) is the \( \Phi2 \) distance between two frequency vectors
The extension of the within-cluster inertia

Let \( D = (d_{ii'}) \) the distance matrix

- the extension of the inertia of a cluster is:

\[
I(C_k) = \sum_{i \in C_k} \sum_{i' \in C_k} \frac{1}{2nn_k} d_{ii'}^2
\]

- the extension of the within-clusters inertia of a partition in \( K \) clusters is:

\[
W(P) = \sum_{k=1}^{K} I(C_k)
\]
The proposed divisive clustering method

Uses a variable and a binary transformation of this variable to split a cluster

- extension of the definition of binary question to the case of symbolic data

Choose the best bipartition among all possible bipartition in 2 clusters

- how many possible bipartition to evaluate?
Binary question and symbolic data

- Tree 1: 
  - C
  - ≤ 172: Taille
    - C₁
  - > 172: C₂
- Tree 2: 
  - C
  - {brun, châtain}
    - C₁
  - {blond, roux}
    - C₂
How many possible bipartitions

If the variable $Y_j$ is quantitative (classical or interval) there is at most $z_j=n_k-1$ bipartitions induced by this variable.

If the variable $Y_j$ is qualitative there is at most:
- $z_j=m_j-1$ bipartitions induced for ordered categories
- $z_j=2^{m_j}-1$ bipartitions induced for non ordered categories

$\Rightarrow z_1+\ldots+ z_p$ bipartitions to evaluate
The proposed divisive clustering method

Choose at each step the cluster to be split such that the new partition has minimum within inertia?

\[ \Delta(C_k) = Q(C_k) - Q(C_k^1) - Q(C_k^2) \]

Stops after L iterations given as input by the user and the output is a hierarchy indexed by \( \Delta \), which is isotonic.
Application for clustering a set of categories

<table>
<thead>
<tr>
<th>FISH</th>
<th>SPECIES</th>
<th>Ln(liver/Muscle)</th>
<th>...</th>
<th>Ln(Stomach/Muscle)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ageneiosus brevifili</td>
<td>-0.12</td>
<td>...</td>
<td>-1.77</td>
</tr>
<tr>
<td></td>
<td>Ageneiosus brevifili</td>
<td>-0.30</td>
<td>...</td>
<td>-1.22</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>Doras micropoeus</td>
<td>0.8</td>
<td>...</td>
<td>-0.89</td>
</tr>
<tr>
<td>67</td>
<td>Doras micropoeus</td>
<td>1.34</td>
<td>...</td>
<td>-1.45</td>
</tr>
</tbody>
</table>

Clustering the 10 categories of SPECIES
First numerical treatment

Clustering of the 67 fish: Ward hierarchical clustering on the 67 fish × 5 concentration variables

<table>
<thead>
<tr>
<th>Fish</th>
<th>Cluster1</th>
<th>Cluster2</th>
<th>Cluster3</th>
<th>Cluster 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ageneiosus brevifili</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cynodon gibbus</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hoplias aimara</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doras micropoeus</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Leporinus fasciatus</td>
<td>33.33</td>
<td>66.67</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Leporinus frederici</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pseudoancistrus barbatus</td>
<td>14.29</td>
<td>0</td>
<td>0</td>
<td>85.71</td>
</tr>
<tr>
<td>Semaprochilodus varii</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Platydoras costatus</td>
<td>20</td>
<td>0</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Potamotrygon hystrix</td>
<td>50</td>
<td>0</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>
First numerical treatment
Second numerical treatment

Clustering of the 10 species on a classical numerical data set…

<table>
<thead>
<tr>
<th>NUM</th>
<th>SPECIES</th>
<th>Ln(liver/Muscle)</th>
<th>...</th>
<th>Ln(Stomach/Muscle)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ageneiosus brevifili</td>
<td>-0.38</td>
<td>...</td>
<td>-1.17</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>10</td>
<td>Doras micropoeus</td>
<td>1.72</td>
<td>...</td>
<td>-0.84</td>
</tr>
</tbody>
</table>
Second numerical treatment

... with the DIV method:

Cluster 1
Ageneiosus brevifilis
Cynodon gibbus
Hoplias aímara
Leporinus frederici

Ln(liver/muscle)

≤ -0.4 > -0.4

Clustering phases:
1. Ln(intestine/muscle)
2. Ln(kidney/muscle)

Cluster 2
Leporinus fasciatus

Cluster 3
Pseudoancistrus barbatus
Doras micropoeus
Potamotrygon hystrix
Platydoras costatus

Cluster 4
Semaprochilodus vari
Doras micropoeus
Potamotrygon hystrix
Platydoras costatus

≤ 1.1 > 1.1

≤ 0.13 > -0.13
Symbolic treatment

Clustering of the 10 species on a symbolic data set...

<table>
<thead>
<tr>
<th>NUM</th>
<th>SPECIES</th>
<th>Ln(liver/Muscle)</th>
<th>…</th>
<th>Ln(Stomach/Muscle)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ageneiosus brevifili</td>
<td>[-0.8;-0.08]</td>
<td>…</td>
<td>[-1.48;-0.59]</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>10</td>
<td>Doras micropoeus</td>
<td>[1.34;2.12]</td>
<td>…</td>
<td>[-1.45;-0.24]</td>
</tr>
</tbody>
</table>
Symbolic treatment

... with the DIV method: