HIERARCHICAL AND PYRAMIDAL CLUSTERING FOR SYMBOLIC DATA

Paula Brito
Univ. Porto, Portugal

OUTLINE

• The hierarchical model
• The pyramidal model
• Numerical hierarchical / pyramidal clustering
• Symbolic Clustering
  • The property of completeness
  • The generality degree
  • The clustering algorithm
• Examples
• The HIPYR Module
Hierarchical Model:
set of nested partitions

Let $E$ be the observations set (the set being clustered)

Hierarchy on $E$:
Family $H$ on non-empty subsets of $E$ such that

- $E \in H$
- $\forall a \in E, \{a\} \in H$
- $\forall h, h' \in H, h \cap h' = \emptyset$ or $h \subseteq h'$ or $h' \subseteq h$
PYRAMIDAL REPRESENTATION


PYRAMID : set of nested overlappings

\[ P = \{ \{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_1, x_2\}, \{x_3, x_4\}, \{x_4, x_5\}, \{x_1, x_2, x_3\}, \{x_1, x_2, x_3, x_4\}, \{x_3, x_4, x_5\}, \{x_1, x_2, x_3, x_4, x_5\} \} \]

PYRAMID P on E

Family on non-empty subsets of E such that:

• \( E \in P \)
• \( \forall a \in E, \{a\} \in P \)
• \( \forall p, p' \in P, \ p \cap p' = \emptyset \) or \( p \cap p' \in P \)
• there exists a linear order \( \theta \) / every element of \( P \) is an interval of \( \theta \)

Pyramidal model:

[Diagram showing Clustering and Seriation]

Hierarchy : nested partitions

Pyramid : nested overlappings
SUCCESSOR AND PREDECESSOR

\( p \) SUCCESSOR of \( p' \) if
- \( p \subseteq p' \)
- \( \exists p'' \iff p \subseteq p'' \subseteq p' \)

\( p' \) is a PREDECESSOR of \( p \)

- Hierarchy: Each cluster has at most ONE predecessor
- Pyramid: Each cluster has at most TWO predecessors

ASCENDING CLUSTERING ALGORITHM

Starting with the one cluster elements, merge at each step the MERGEABLE clusters for which the dissimilarity (aggregation index) is MINIMUM

MERGEABLE CLUSTERS:

- \( \rightarrow \) if the structure is a hierarchy: none of them has been aggregated before;

- \( \rightarrow \) if the structure is a pyramid: none of them has been aggregated twice, and \( p \) is an interval of a total order \( \emptyset \) on \( E \)
SYMBOLIC CLUSTERING

OBJECTIVE:
Given a set of symbolic objects, build an hierarchical / pyramidal clustering such that each cluster is a pair

- EXTENSION - its members
- INTENSION - its description

EACH CLUSTER HAS AN AUTOMATIC SYMBOLIC REPRESENTATION
BY A SYMBOLIC OBJECT

Symbolic clustering methods need:

- Generalization Operator
  \[ C \subseteq C' \]
  \[ s' \text{ (representing } C') \text{ is more general than } s \text{ (representing } C) \]
- Generality degree measure
GENERALISATION

s is more general than s’ if its extent contains the extent of s’
s’ is more specific than s

Generalisation of two symbolic objects s and s’:
determining s’’ : s’’ is more general than both s and s’.

\[ s \subseteq s \cup s' \quad \text{and} \quad s' \subseteq s \cup s' \]
\[ \text{ext} (s \cup s') \supseteq \text{ext} (s) \quad \text{and} \quad \text{ext} (s \cup s') \supseteq \text{ext} (s') \]

This procedure differs according to the variable type :

1) Interval variables

\[ s_1 = [ y \in [a_1, b_1] ] \]
\[ s_2 = [ y \in [a_2, b_2] ] \]
\[ s_1 \cup s_2 = [ y \in [\min \{a_1,a_2\}, \max \{b_1,b_2\}]] \]

Example :

\[ s_1 = [ \text{time} \in [5, 15] ] \]
\[ s_2 = [ \text{time} \in [10, 20] ] \]
\[ s_1 \cup s_2 = [ \text{time} \in [5, 20]] \]
2) Multi-valued categorical variables

\[ s_1 = [y \in V1] \]
\[ s_2 = [y \in V2] \]
\[ s_1 \cup s_2 = [y \in V1 \cup V2] \]

Example:

\[ s_1 = \text{[job} \in \{\text{secretary, teacher}\}\text{]} \]
\[ s_2 = \text{[job} \in \{\text{employee}\}\text{]} \]
\[ s_1 \cup s_2 = \text{[job} \in \{\text{secretary, teacher, employee}\}\text{]} \]

3) Modal variables

Two possibilities proposed:

- take for each category the Maximum of its frequencies
- take for each category the Minimum of its frequencies
a) Generalisation by the Maximum

\[ y = \{ m_1(p_1^1), \ldots, m_k(p_k^1) \} \cup \{ m_1(p_1^2), \ldots, m_k(p_k^2) \} = \]

\[ y = \{ m_1(p_1), \ldots, m_k(p_k) \} \]

with \( p_j = \text{Max} \{ p_j^1, p_j^2 \} \)

Example:

\[ \text{Type of Job } \in \{ (0.3) \text{ administration, (0.7) teaching} \} \cup \]

\[ \text{Type of Job } \in \{ (0.6) \text{ admin., (0.2) teaching, (0.2) secretary} \} \]

\[ = \text{Type of Job } \in \{ (0.6) \text{ admin., (0.7) teaching, (0.2) secret.} \} \]

Extension: \( \{ a : p_j^a \leq p_j^1, j = 1, \ldots, k \} \)

\[ \rightarrow \text{ "at most" principle} \]

b) Generalisation by the Minimum

\[ y = \{ m_1(p_1^1), \ldots, m_k(p_k^1) \} \cup \{ m_1(p_1^2), \ldots, m_k(p_k^2) \} = \]

\[ y = \{ m_1(p_1), \ldots, m_k(p_k) \} \]

with \( p_j = \text{Min} \{ p_j^1, p_j^2 \} \)

Example:

\[ \text{Type of Job } \in \{ (0.3) \text{ administration, (0.7) teaching} \} \cup \]

\[ \text{Type of Job } \in \{ (0.6) \text{ admin., (0.2) teaching, (0.2) secretary} \} \]

\[ = \text{Type of Job } \in \{ (0.3) \text{ admin., (0.2) teaching} \} \]

Extension: \( \{ a : p_j^a \geq p_j, j = 1, \ldots, k \} \)

\[ \rightarrow \text{ "at least" principle} \]
COMPLETENESS AND COMPLETE OBJECTS

COMPLETE symbolic object

- defined by all the properties that characterize its extension
- the most specific to fulfill this condition

→ Galois connections, lattice theory
→ Barbut & Monjardet (1970); Wille (1982); Ganter (1984)

Union preserves completeness

Example:

<table>
<thead>
<tr>
<th></th>
<th>y_1</th>
<th>y_2</th>
<th>y_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>20</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>55</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>50</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>60</td>
<td>M</td>
</tr>
</tbody>
</table>

\[ a = \left[ \text{Weight} = [40, 50]\right] \land \left[ \text{Age} = [20, 50]\right] \]
\[ \text{ext } a = \{\text{w1, w3}\} \]
\[ a \text{ IS NOT COMPLETE} \]

\[ a = \left[ \text{Weight} = [45, 50]\right] \land \left[ \text{Age} = [20, 30]\right] \land \left[ \text{Sex} = \{\text{F}\}\right] \]
\[ \text{IS COMPLETE} \]
THE BASIC ALGORITHM

STARTING WITH THE ONE-OBJECT CLUSTERS
\{a_i\}, i = 1,\ldots,n

At each step, FORM A CLUSTER \( p \) union of \( p_1, p_2 \), REPRESENTED BY \( s \) such that

- \( p_1, p_2 \) can be merged together
- \( s \) is complete
- \( s \) is more general than \( s_1, s_2 \Rightarrow \text{union } s_1 \cup s_2 \)
- \( \text{ext}_E s = p \)

\[ \Rightarrow \quad \text{Non-uniqueness } \Rightarrow \text{numerical criterion} \quad \text{Clusters with more specific descriptions are formed first} \]

GENERALITY DEGREE

\[ a = \wedge \{ y_i \in V_i \}, \quad V_i \subseteq O_i \quad O_i \text{ bounded} \]

\[ G(a) = \prod_{i=1}^{p} \frac{m(V_i)}{m(O_i)} = \prod_{i=1}^{p} G(e_i) \]

if \( a = \wedge e_i \)

PROPORTION of the description space covered by \( a \)

\[ \Rightarrow \quad \text{the more possible members of the extension of } a, \quad \text{the greater the generality degree of } a \]
Interval variables

\[ m(V_i) = \max V_i - \min V_i \quad \text{(range)} \]

Example:
Describing groups of people defined on variables age and salary.
Age ranges from 15 to 60, salary ranges from 0 to 10000.

Consider a group described by a symbolic object
\[ s_1 = [\text{age } \in [20, 45]] \land [\text{salary } \in [1000, 3000]] = e_{11} \land e_{12} \]

\[
G(e_{11}) = \frac{45 - 20}{60 - 15} = \frac{25}{45} = 0.55
\]

\[
G(e_{12}) = \frac{3000 - 1000}{10000 - 0} = \frac{2000}{10000} = 0.2
\]

\[ G(s_1) = 0.55 \ast 0.2 = 0.11 \]

Multi-valued categorical variables

\[ m(V_i) = \text{card } V_i \]

Example:
Describing groups of people from the UE, defined on variables sex and nationality.

Consider one group described by:
\[ s_1 = [\text{sex } \in \{M\}] \land [\text{nationality } \in \{\text{French, English}\}] \]

\[
G(e_{11}) = \frac{1}{2} = 0.5 \\
G(e_{12}) = \frac{2}{25} = 0.08
\]

\[ G(s_1) = 0.5 \times 0.08 = 0.04 \]
Modal variables

a) Generalising by the Maximum

\[ G_1(a) = \prod_{j=1}^{p} \frac{1}{\sqrt{k_j}} \sum_{i=1}^{k_j} \sqrt{p_{ij}} \]

which is the **affinity coefficient** (Matusita, 1951) between \((p_1, \ldots, p_k)\) and the uniform distribution

\[ G_1(a) \text{ is maximum (=1) when } p_i = 1/k, i = 1, \ldots, k : \text{uniform} \]

This means that we consider an object
the more general the more similar it is
to the uniform distribution

b) Generalising by the minimum

\[ G_2(a) = \prod_{j=1}^{p} \frac{1}{\sqrt{k_j(k_j-1)}} \sum_{i=1}^{k_j} \sqrt{(1 - p_{ij})} \]

Again, \( G_2(a) \) is maximum (=1)

when \( p_i = 1/k, i = 1, \ldots, k : \text{uniform} \)
ALGORITHM

Starting with the one-object clusters \( \{a_i\}, i = 1, \ldots, n \)

At each step, FORM A CLUSTER \( p \) union of \( p_1, p_2 \), 
REPRESENTED BY \( s \) such that

- \( p_1, p_2 \) can be merged together
- \( s = s_1 \cup s_2 \) (complete)
- \( \text{ext}_E s = p \)
- \( G(s) \) is minimum

The algorithm builds a hierarchy / pyramid on \( E \), such that

each cluster is represented by a "complete" symbolic object
whose extension is the cluster itself

\[ \text{CLUSTER} = (p, s) \quad p = \text{ext} s \]

\[ \rightarrow \text{AUTOMATIC SYMBOLIC REPRESENTATION OF THE CLUSTERS} \]
\[
\begin{array}{cccc}
 y_1 & y_2 & y_3 & y_4 \\
 w_1 & 1 & 1 & 1 & 2 \\
 w_2 & 1 & 2 & 1 & 3 \\
 w_3 & 1 & 2 & 2 & 2 \\
 w_4 & 2 & 1 & 1 & 2 \\
 w_5 & 3 & 3 & 2 & 1 \\
\end{array}
\]

\[ P_6 : s_6 = [ y_1 = \{1\} ] \land [ y_2 = \{1,2\} ] \land [ y_4 = \{2,3\} ] \]

\[ P_7 : s_7 = [ y_1 = \{1,2\} ] \land [ y_2 = \{1,2\} ] \land [ y_4 = \{2,3\} ] \]
Application

INE Labour Force Survey data:

12 sex * age groups

Each of the 12 groups is described by interval variables, representing the variation of the underlying variables in the class.

<table>
<thead>
<tr>
<th>sex / age groups</th>
<th>Full-time</th>
<th>Part-time</th>
<th>Primary studies</th>
<th>Secondary studies</th>
<th>University study</th>
<th>Without study</th>
<th>Employees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men / 15 to 24</td>
<td>[54.50, 66.50]</td>
<td>[3.50, 5.50]</td>
<td>[54.50, 56.00]</td>
<td>[36.50, 41.00]</td>
<td>[1.10, 2.70]</td>
<td>[1.00, 3.20]</td>
<td>[66.50, 69.00]</td>
</tr>
<tr>
<td>Men / 25 to 34</td>
<td>[56.50, 67.00]</td>
<td>[2.40, 3.70]</td>
<td>[57.50, 61.10]</td>
<td>[27.60, 31.00]</td>
<td>[6.00, 8.00]</td>
<td>[2.90, 4.30]</td>
<td>[69.20, 73.10]</td>
</tr>
<tr>
<td>Men / 35 to 44</td>
<td>[67.50, 68.80]</td>
<td>[1.20, 2.10]</td>
<td>[60.70, 64.20]</td>
<td>[23.00, 27.00]</td>
<td>[6.70, 8.90]</td>
<td>[3.70, 5.20]</td>
<td>[71.30, 74.50]</td>
</tr>
<tr>
<td>Men / 45 to 54</td>
<td>[58.60, 67.00]</td>
<td>[2.10, 3.40]</td>
<td>[57.10, 70.00]</td>
<td>[16.60, 19.50]</td>
<td>[6.80, 9.00]</td>
<td>[4.30, 5.10]</td>
<td>[65.20, 69.20]</td>
</tr>
<tr>
<td>Women / 15 to 24</td>
<td>[67.20, 70.10]</td>
<td>[9.80, 12.80]</td>
<td>[63.30, 72.70]</td>
<td>[7.10, 9.70]</td>
<td>[6.80, 9.20]</td>
<td>[6.80, 9.50]</td>
<td>[67.20, 72.50]</td>
</tr>
<tr>
<td>Men / 25 to 34</td>
<td>[53.20, 59.10]</td>
<td>[4.20, 4.80]</td>
<td>[45.50, 51.70]</td>
<td>[4.20, 7.30]</td>
<td>[3.90, 5.80]</td>
<td>[3.80, 4.80]</td>
<td>[52.60, 57.30]</td>
</tr>
<tr>
<td>Women / 35 to 44</td>
<td>[66.20, 69.70]</td>
<td>[9.30, 11.90]</td>
<td>[48.10, 53.20]</td>
<td>[14.00, 17.00]</td>
<td>[1.10, 2.40]</td>
<td>[62.50, 65.90]</td>
<td></td>
</tr>
<tr>
<td>Women / 45 to 54</td>
<td>[67.40, 68.90]</td>
<td>[10.10, 12.00]</td>
<td>[50.90, 60.60]</td>
<td>[22.05, 25.40]</td>
<td>[12.16, 14.90]</td>
<td>[2.10, 4.70]</td>
<td>[74.00, 77.90]</td>
</tr>
<tr>
<td>Women / 55 to 64</td>
<td>[10.00, 55.40]</td>
<td>[16.90, 20.00]</td>
<td>[61.80, 85.69]</td>
<td>[12.10, 15.10]</td>
<td>[8.00, 11.70]</td>
<td>[10.90, 13.60]</td>
<td>[65.30, 68.40]</td>
</tr>
<tr>
<td>Women / 65 to 74</td>
<td>[61.50, 67.00]</td>
<td>[13.00, 18.40]</td>
<td>[43.80, 49.10]</td>
<td>[2.60, 5.10]</td>
<td>[4.00, 7.20]</td>
<td>[41.00, 45.60]</td>
<td>[41.10, 46.00]</td>
</tr>
<tr>
<td>Women / 75 and over</td>
<td>[65.30, 68.00]</td>
<td>[55.70, 64.70]</td>
<td>[22.20, 29.30]</td>
<td>[0.80, 3.00]</td>
<td>[1.30, 3.80]</td>
<td>[68.10, 73.60]</td>
<td>[12.70, 19.80]</td>
</tr>
</tbody>
</table>
P36


[Financial_Act.=[3.639, 5.851]]

[Public_Admin=[1.336, 2.620]U[3.004, 4.871]]

[Hotels_Rest.=[1.874, 3.497]]

[Commerce=[9.582, 13.161]]

[Construction=[11.244, 14.526]]

[Education=[4.589, 8.829]]

[Elect_gas_water=[0.347, 1.757]]

[Industry=[32.700, 37.277]U[38.671, 43.205]]

[Other_Serv.=[4.316, 7.551]]

[Primary=[5.921, 9.799]]

[Health=[1.857, 3.622]]

[Transp_Comunic.=[2.352, 4.675]]

its members

Man 15 to 54
Women 15 to 24

are the only elements fulfilling these conditions.

Application

Cultural activities of 11 socio-professional groups
1509 individual observations grouped
<table>
<thead>
<tr>
<th>Category</th>
<th>Cinema</th>
<th>Football</th>
<th>Daily News</th>
<th>Weekly Nature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>no (0.40), yes (0.60)</td>
<td>no (0.49), yes (0.51), no_an (0.31)</td>
<td>yes (0.80), no (0.39), no_an (0.41)</td>
<td>no (0.73), yes (0.27)</td>
</tr>
<tr>
<td>Retired</td>
<td>no (0.62), yes (0.38), no_an (0.30)</td>
<td>no (0.88), yes (0.12)</td>
<td>yes (0.44), no (0.56)</td>
<td>no (0.90), yes (0.09)</td>
</tr>
<tr>
<td>Employed</td>
<td>yes (0.72), yes (0.28), no_an (0.00)</td>
<td>no (0.73), yes (0.22)</td>
<td>yes (0.97), no (0.03)</td>
<td>no (0.93), yes (0.07)</td>
</tr>
<tr>
<td>Small Index</td>
<td>no (0.80), yes (0.20)</td>
<td>no (0.67), yes (0.32), no_an (0.01)</td>
<td>yes (0.71), no (0.28), no_an (0.01)</td>
<td>no (0.68), yes (0.32)</td>
</tr>
<tr>
<td>Housewife</td>
<td>no (0.91), yes (0.09)</td>
<td>no (0.60), yes (0.40)</td>
<td>yes (0.30), no (0.70)</td>
<td>no (0.62), yes (0.38)</td>
</tr>
<tr>
<td>Medium Staff</td>
<td>no (0.62), yes (0.38), no_an (0.01)</td>
<td>no (0.73), yes (0.25), no_an (0.01)</td>
<td>yes (0.79), no (0.21)</td>
<td>no (0.88), yes (0.12)</td>
</tr>
<tr>
<td>Ind. Worker</td>
<td>no (0.72), yes (0.27), no_an (0.01)</td>
<td>no (0.60), yes (0.40)</td>
<td>yes (0.80), no (0.40)</td>
<td>no (0.68), yes (0.32)</td>
</tr>
<tr>
<td>Other Staff</td>
<td>no (0.44), yes (0.56)</td>
<td>no (0.78), yes (0.22)</td>
<td>yes (0.94), no (0.16)</td>
<td>no (0.20), yes (0.80)</td>
</tr>
<tr>
<td>Other</td>
<td>no (0.78), yes (0.22), no_an (0.01)</td>
<td>no (0.60), yes (0.19), no_an (0.01)</td>
<td>yes (0.48), no (0.52)</td>
<td>no (0.87), yes (0.13)</td>
</tr>
<tr>
<td>Dir./Lib.</td>
<td>no (0.49), yes (0.61)</td>
<td>no (0.70), yes (0.29), no_an (0.02)</td>
<td>yes (0.88), no (0.12)</td>
<td>no (0.85), yes (0.15)</td>
</tr>
<tr>
<td>Manager</td>
<td>no (0.42), yes (0.58)</td>
<td>no (0.50), yes (0.42)</td>
<td>yes (0.75), no (0.25)</td>
<td>no (0.50), yes (0.50)</td>
</tr>
</tbody>
</table>
The HIPYR Module

Objective:

Perform Hierarchical or Pyramidal clustering on a set of SO’s

✓ from a dissimilarity matrix → numerical clustering

✓ directly based on the data set → symbolic clustering: clusters are "concepts"

The HIPYR Module
Main Parameters

Structure: Hierarchy or Pyramid

Data Source:
- Dissimilarity Matrix (Numerical Clustering)
- Symbolic objects (Symbolic Clustering)

Aggregation Index:
- Numerical Clustering: Maximum, Minimum, Average, Diameter
- Symbolic Clustering: Minimum Generality, Minimum Increase in Generality

Main Parameters

- Order Variable (optional): quantitative single variable; to impose an order compatible with the pyramid
- Modal variables generalization:
  - Maximum
  - Minimum
- Use Taxonomies for generalization (nominal or categorical multi-valued variables): Y, N
- Select “best” classes: Y, N
- Write induced dissimilarity/generality matrix: Y, N
Induced dissimilarity/generality matrix

For each pair of SO, $s_i$, $s_j$, $i, j, =1, ..., n$,

\[ d^*(s_i, s_j) = \text{index (height) of the “smallest” class that contains } s_i \text{ and } s_j \]

\[ d^*(s_i, s_j) = \text{Min } \{ f(C), s_i \in C, s_j \in C \} \]

Evaluation of the obtained indexed hierarchy / pyramid:
Comparison between the initial and the induced dissimilarity/generality matrices.

Evaluation value

For $s_i$, $s_j$, $i, j, =1, ..., n$, $d(s_i, s_j)$:
- the given dissimilarity matrix (numerical clustering)
- generality degree of $s_i \cup s_j$ (symbolic clustering)

\[
EV = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (d(s_i, s_j) - d^*(s_i, s_j))^2}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d(s_i, s_j)}
\]
Cluster selection

Identify the most interesting clusters:

A cluster is “interesting” if its variability is small as compared to its predecessors.

Variability indicated by index values f(h).

Compute mean value and standard deviation of height increase values.

A class is selected if the corresponding increase value is more than 2 stand. dev. over the mean value.
HIPYR

OUTPUT

- Text file
- Sodas file
- Interactive Graphical Representation (VPYR)

The output listing contains:

- The labels of the individuals
- The labels of the variables
- The description of each node:
  - the symbolic object associated to each node
  - its extension
- Evaluation value
- Selected clusters, if asked for
- The induced matrix, if asked for
Options of the Graphical Representation

A cluster is selected by clicking on it.

- Description of the cluster in terms of list of chosen variables
- Representation by a Zoom Star
Options of the Graphical Representation

Pruning the hierarchy or pyramid using the aggregation heights as a criterion.
Suppressing cluster p if:
\[ f(p') - f(p) < \alpha f(E) \land p \text{ has a single predecessor} \]

Rate of simplification \( \alpha \) chosen by the user,
new graphic window with the simplified structure.
Rule Generation

If the hierarchy/pyramid are built from a symbolic data table, rules may be generated and saved in a specified file.

Fission method:
\[ s \Rightarrow s_1 \lor s_2 \]

Fussion method (pyramids only):
\[ s_1 \land s_2 \Rightarrow s \]
Rule generation

Options of the Graphical Representation

Reduction

Should the user be interested in a particular cluster, he may obtain a window with the structure restricted to this cluster and its successors.
Reduction