Clustering Symbolic Data

Paula Brito

Fac. Economia & LIAAD-INESC TEC, Univ. Porto, Portugal

ECI 2015 - Buenos Aires
T3: Symbolic Data Analysis:
Taking Variability in Data into Account
Outline

1. Clustering approaches

2. Divisive Clustering
   - The criterion
   - The distances
   - Binary questions and Assignement
   - The algorithm

3. Hierarchical and pyramidal (conceptual) clustering
   - Generality Measure

4. Non-Hierarchical clustering
Outline

1. Clustering approaches

   Divisive Clustering
   - The criterion
   - The distances
   - Binary questions and Assignment
   - The algorithm

2. Hierarchical and pyramidal (conceptual) clustering
   - Generality Measure

3. Non-Hierarchical clustering
Clustering interval-valued and distributional data

→ Necessary to define, or adapt, clustering methods

Two groups of methods:

a) Methods based on Dissimilarities: adapting classical clustering methods to the new kind of data

b) Conceptual clustering methods: use the data explicitly in the clustering process, classes usually described by necessary and sufficient conditions based on Generalization procedures
Clustering interval and distributional data

Type a) : require appropriate dissimilarity measures

- Many measures proposed in the literature

- Interval-valued data:
  - Minkowski-type distances
  - Malahanobis distance
  - Hausdorff distance

- Distribution-valued data
  - Wasserstein distance
  - Mallows distance
Clustering approaches

Divisive Clustering
Hierarchical and pyramidal (conceptual) clustering
Non-Hierarchical clustering

Clustering interval-valued data

- *K*-means-like approaches - De Carvalho and co-workers (2004 - ...)
  - Different distances considered
  - Also: Adaptive distances
  - Also: Multiple dissimilarity matrices
  - Using Hausdorff distance - Chavent & Lechevallier (2002)
- Fuzzy clustering
  - El-Sonbaty, Ismail (1998)
  - Yang, Hwang and Chen (2004)
  - D’Urso and Giordani (2006)
  - De Carvalho *et al* (2007, 2010)
  - Jeng, Chuang, Tseng and Juan (2010)
- SOM approaches:
  - De Carvalho *et al* (2011)
  - Hajjar and Hamdan (2011)
  - Yang, Hung, Chen (2012)
Interval-valued variables: Distance measures

Many measures proposed in the literature

**Hausdorff distance** :
\[ d_H(l_i, l_j) = \max \{\{|l_i - l_j|, |u_i - u_j|\} \}

**Euclidean distance** :
\[ d_2(l_i, l_j) = \sqrt{(l_i - l_j)^2 + (u_i - u_j)^2} \]

**City-Block distance** :
\[ d_1(l_i, l_j) = |l_i - l_j| + |u_i - u_j| \]

**Mahalanobis distance**:
defined on the basis of the vectors of observed lower \( X_{iL} = (l_{i1}, \ldots, l_{ip}) \) and upper bounds \( X_{iU} = (u_{i1}, \ldots, u_{ip}) \).
\[ d(s_{i_1}, s_{i_2}) = d_M(X_{i_1L}, X_{i_2L}) + d_M(X_{i_1U}, X_{i_2U}) \]
where
\[ d_M(X_{i_1L}, X_{i_2L}) = (X_{i_1L} - X_{i_2L})^t M_L(X_{i_1L} - X_{i_2L}) \]
is the Mahalanobis distance between the two vectors \( X_{i_1L} \) and \( X_{i_2L} \)
\[ d_M(X_{i_1U}, X_{i_2U}) = (X_{i_1U} - X_{i_2U})^t M_U(X_{i_1U} - X_{i_2U}) \]
Clustering distributional data

- Hardy (2004, 2008) developed SHICLUST
  - extends single and complete linkage, centroid and Ward methods to categorical modal variables
  - dissimilarity measures and aggregation indices adapted or suitably chosen

  - used the Mallows distance for clustering histogram-valued data
  - rewrote it using the centre and half-range of the subintervals
  - both hierarchical and dynamical clustering approaches

- Korenjak-Cerne et al (2011)
  - two clustering methods for data with discrete distributions
  - the adapted leaders method and the adapted Ward’s method
  - descriptions with distributions allow combining two separate data sets into a single one
### Clustering approaches
- Divisive Clustering
- Hierarchical and pyramidal (conceptual) clustering
- Non-Hierarchical clustering

### Histogram-valued variables: Distance measures

Many measures proposed in the literature (see e.g. Bock and Diday (2000), Gibbs, (2002))

<table>
<thead>
<tr>
<th>Divergence measures</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kullback-Leibler</td>
<td>$D_{KL}(f, g) = \int_{\mathbb{R}} \log \left( \frac{f(x)}{g(x)} \right) f(x) dx$</td>
</tr>
<tr>
<td>Jeffrey</td>
<td>$D_J(f, g) = D_{KL}(f, g) + D_{KL}(g, f)$</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>$D_{\chi^2}(f, g) = \int_{\mathbb{R}} \frac{</td>
</tr>
<tr>
<td>Hellinger</td>
<td>$D_H(f, g) = \left[ \int_{\mathbb{R}} \left( \sqrt{f(x)} - \sqrt{g(x)} \right) dx \right]^{1/2}$</td>
</tr>
<tr>
<td>Total variation</td>
<td>$D_{var}(f, g) = \int_{\mathbb{R}}</td>
</tr>
<tr>
<td>Wasserstein</td>
<td>$D_W(f, g) = \int_{\mathbb{R}}</td>
</tr>
<tr>
<td>Mallows</td>
<td>$D_M(f, g) = \sqrt{\int_0^1 (F^{-1}(x) - G^{-1}(x))^2 dx}$</td>
</tr>
<tr>
<td>Kolmogorov</td>
<td>$D_W(f, g) = \max_{\mathbb{R}}</td>
</tr>
</tbody>
</table>
Outline

1. Clustering approaches
2. Divisive Clustering
   - The criterion
   - The distances
   - Binary questions and Assignment
   - The algorithm
3. Hierarchical and pyramidal (conceptual) clustering
   - Generality Measure
4. Non-Hierarchical clustering
DIV: Divisive Clustering (Chavent (1998, 2000); Brito, Chavent (2012))

- Divisive clustering method
- For symbolic data
- Taking internal variability into account
- Monothetic clusters

- In SODAS: Interval and Categorical modal variables (not mixed)
- More recently: Method for Interval and Histogram-valued variables
- Where: Interval-valued variables: a special case of histogram-valued variables
Divisive Clustering

- Divisive clustering algorithms proceed top-down
- Starting with \( S \), the set to be clustered
- Performing a bipartition of one cluster at each step
- At step \( m \) a partition of \( S \) in \( m \) clusters is present
- One will be further divided in two sub-clusters
- The cluster to be divided and the splitting rule chosen to obtain a partition in \( m + 1 \) clusters minimizing intra-cluster dispersion
The criterion

“Quality” of a partition \( P_m = \{ C_1^{(m)}, C_2^{(m)}, \ldots, C_m^{(m)} \} \) measured by the sum of intra-cluster dispersion for each cluster:

\[
Q(m) = \sum_{\alpha=1}^{K} I(C_{\alpha}) = \sum_{\alpha=1}^{K} \sum_{s_i, s_i' \in C_{\alpha}^{(m)}} D^2(s_i, s_i')
\]

\[
D^2(s_i, s_i') = \sum_{j=1}^{p} d^2(x_{ij}, x_{i'j})
\]

\( d \) : quadratic distance between distributions

At each step:

one cluster is chosen to be split in two sub-clusters

\( Q(m + 1) \) is minimized (\( Q(m) - Q(m + 1) \) maximized)
Distance measures: Interval data in SODAS

Hausdorff distance

The Hausdorff distance between two sets is the maximum distance of a set to the nearest point in the other set. Two sets are close if every point of either set is close to some point of the other set.

Hausdorff distance between two intervals $I_1 = [l_1, u_1], I_2 = [l_2, u_2]$:

$$d_H(I_1, I_2) = \max\{|l_1 - l_2|, |u_1 - u_2|\}$$

For multivariate interval-valued observations these may be combined, often in an “Euclidean” way:

$$d_{H_2}(s_{i1}, s_{i2}) = \sqrt{\sum_{j=1}^{p} (\max\{|l_{i1j} - l_{i2j}|, |u_{i1j} - u_{i2j}|\})^2}$$
Distance measures: Categorical modal data in SODAS

The symbolic data array is transformed in a frequency matrix
\( X = (f_{kj})_{nt} \) with \( t = \) total number of categories

Let \( p_{kj} = \frac{f_{kj}}{np} \)

\( \chi^2 \) distance:

\[
d(s_k, s_\ell) = \sum_{j=1}^{t} \frac{p_.}{p.j} \left( \frac{p_{kj}}{p_.} - \frac{p_{\ell j}}{p_.} \right)^2
\]
Distance measures: New approach

Evaluate the dissimilarity between distributions

\[ Y_j(S_i) = H_{Y_j(S_i)} = ([l_{ij1}, \bar{l}_{ij1}, p_{ij1}; \ldots; [l_{ijK_j}, \bar{l}_{ijK_j}], p_{ijK_j}) \]

1. **Mallows distance**

\[ d^2_M(x_{ij}, x'_{ij}) = \int_0^1 (q_{ij}(t) - q'_{ij}(t))^2 dt \]

\( q_{ij} \) : quantile function corresponding to distribution \( Y_j(S_i) \)

2. **Squared Euclidean distance**

\[ d^2_E(x_{ij}, x'_{ij}) = \sum_{\ell=1}^{K_j} (p_{ij\ell} - p'_{ij\ell})^2 \]
Binary questions

Bipartition to be performed at each step
defined by one single variable considering conditions of the type :

- Numerical data :
  \[ R_{j\ell} := Y_j \in R_{1j} \iff Y_j \leq m_\ell, j = 1, \ldots, p \]

- Categorical data :
  \[ R_{j\ell} := Y_j \in R_{1j} \]

\( R_{j\ell} \rightarrow \) bipartition of a cluster :

sub-cluster 1 : elements who verify condition \( R_{j\ell} \) : \( Y_j \in R_{1j} \)
sub-cluster 2 : those who do not : \( Y_j \notin R_{1j} \)
Bi-partitions and Assignment

- Interval variables in SODAS:
  - $m_\ell$ - defining the “cuts” are the midpoints between the centres of the observed intervals
  - Test made with the observed centres

- Categorical modal variables in SODAS
  - Cuts defined by all bi-partitions of the set of categories:
  - $R_{j\ell}$ : sum of category weights in $Y_j(s_i) \geq 0.5$

- New approach for distributional variables:
  - $m_\ell$ - defining the “cuts” are the $I_{j\ell}$
  - $R_{j\ell}$ : $Y_j \leq I_{j\ell}$ iff $\sum_{\alpha=1}^{\ell} p_{ij\alpha} \geq 0.5$

The sequence of conditions:
necessary and sufficient condition for cluster membership

The obtained clustering is monothetic:
each cluster is represented by a conjunction of properties in the descriptive variables
Binary questions and assignment: example

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>([10, 11], 0.50; [11, 12], 0.50; [12, 14], 0.00)</td>
<td>([5, 10], 0.2; [10, 12], 0.5; [12, 14], 0.133; [14, 15], 0.067; [15, 16], 0.033; [16, 18], 0.067; [18, 19], 0.0)</td>
</tr>
<tr>
<td>Class 2</td>
<td>([10, 11], 0.00; [11, 12], 0.33; [12, 14], 0.67)</td>
<td>([5, 10], 0.05; [10, 12], 0.3; [12, 14], 0.25; [14, 15], 0.1; [15, 16], 0.1; [16, 18], 0.133; [18, 19], 0.067)</td>
</tr>
</tbody>
</table>

First step - binary questions:

Age ≤ 11, Age ≤ 12
Marks ≤ 10, Marks ≤ 12, Marks ≤ 14, Marks ≤ 15, Marks ≤ 16, Marks ≤ 18

If condition Age ≤ 12 is selected:
sub-cluster 1 contains Class 1 and is described by “Age ≤ 12”
sub-cluster 2 contains Class 2 and is described by “Age > 12”
Algorithm

- Initialization: \( P_1 = \{ C_1^{(1)} \equiv S \} \)
- \( P_m = \{ C_1^{(m)}, \ldots, C_m^{(m)} \} \): current partition at step \( m \)

Determine the cluster \( C_M^{(m)} \) and the binary question \( R_{j\ell} := Y_j(s_i) \in R_{1j} \):

new partition \( P_{m+1} = \{ C_1^{(m+1)}, \ldots, C_{m+1}^{(m+1)} \} \) minimizes

\[
Q(m) = \sum_{\ell=1}^{m} \sum_{s_i, s_i' \in C_\ell^{(m)}} D^2(s_i, s_i')
\]

among partitions in \( m + 1 \) clusters obtained by splitting a cluster of \( P_m \) in two clusters

- Minimize \( Q(m) \): equivalent to maximize

\[
\Delta Q = I(C_M^{(m)}) - (I(C_1^{(m+1)}) + I(C_2^{(m+1)}))
\]

- Fixed number of clusters \( K \) is attained

or \( P \) has \( n \) clusters, each with a single element (step \( n \):

algorithm stops
Divisive Clustering: Application

Price and Engine Displacement ($cm^3$) of utilitarian cars’ models

<table>
<thead>
<tr>
<th></th>
<th>Price</th>
<th>Engine Displacement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>([15, 25[, 0.5; [25, 35[, 0.5);</td>
<td>([1300, 1500[, 0.2; [1500, 1700[, 0.5; [1700, 1900[, 0.3)</td>
</tr>
<tr>
<td>Model 2</td>
<td>([15, 25[, 0.2; [25, 35[, 0.8);</td>
<td>([1300, 1500[, 0.1; [1500, 1700[, 0.2; [1700, 1900[, 0.7)</td>
</tr>
<tr>
<td>Model 3</td>
<td>([15, 25[, 0, 33; [25, 35[, 0.67)</td>
<td>([1300, 1500[, 0.1; [1500, 1700[, 0.4; [1700, 1900[, 0.5)</td>
</tr>
<tr>
<td>Model 4</td>
<td>([15, 25[, 0.6; [25, 35[, 0.4)</td>
<td>([1300, 1500[, 0.6; [1500, 1700[, 0.4; [1700, 1900[, 0.0)</td>
</tr>
</tbody>
</table>

- Partition into three clusters
- Squared Euclidean distance between distributions to compare the observed values for each car model
Divisive clustering application: the clustering tree

Cluster $C_1^{(3)} = \{\text{Model 4}\}$: “Price $\leq 25 \land \text{Engine Displacement} \leq 1500$”

Cluster $C_2^{(3)} = \{\text{Model 1}\}$: “Price $\leq 25 \land \text{Engine Displacement} > 1500$”

Cluster $C_3^{(3)} = \{\text{Model 2, Model 3}\}$: “Price $> 25$”
Car example: the dendrogram
Application: Social and crime data in USA states

- Data gathered for 2216 USA cities, aggregated by state - 22 states retained
- 14 numerical variables - distributions represented by histogram-valued variables
- Partition into six clusters
- Mallows distance between distributions for each state
Crime-data application: the variables

CRIMES
- murdPerPop: number of murders per 100K population
- robbPerPop: number of robberies per 100K population
- assaultPerPop: number of assaults per 100K population
- burglPerPop: number of burglaries per 100K population
- larcPerPop: number of larcenies per 100K population
- autoTheftPerPop: number of auto thefts per 100K population
- arsonsPerPop: number of arsons per 100K population

SOCIAL
- perCapInc: per capita income
- PctPopUnderPov: percentage of people under the poverty level
- PersPerOccupHous: mean persons per household
- PctKids2Par: percentage of kids in family housing with two parents
- PctVacantBoarded: percent of vacant housing that is boarded up
- NumKindsDrugsSeiz: number of different kinds of drugs seized
- LemasTotReqPerPop: total requests for police per 100K population
Clustering approaches
Hierarchical and pyramidal (conceptual) clustering
Non-Hierarchical clustering

Divisive Clustering

The criterion
The distances
Binary questions and Assignment
The algorithm

Crime-data application: the dendrogram

Outline

1. Clustering approaches

2. Divisive Clustering
   - The criterion
   - The distances
   - Binary questions and Assignment
   - The algorithm

3. Hierarchical and pyramidal (conceptual) clustering
   - Generality Measure

4. Non-Hierarchical clustering
Hierarchical/Pyramidal Conceptual clustering method (Brito (1991, 1995))

- Ascending hierarchical / pyramidal clustering
- Each cluster formed is associated with a conjunction of properties in the input variables
  Cluster = Concept : (extent, description)
- When two given clusters are merged
  - Set-valued and Interval-valued variables: Generalization is performed by union, e.g.:
    \[[0, 15] \cup [10, 30] = [0, 30] \]
  - Distribution-valued variables: Generalization is performed by either considering the maximum or the minimum of the probability/frequency values for each category, e.g.
    \(([[0, 15], 0, 5; [15, 30], 0, 5]) \cup_{\text{max}} ([0, 15], 0, 2; [15, 30], 0, 8) = ([0, 15], 0, 5; [15, 30], 0, 8)\)
- Only clusters corresponding to concepts are formed: the cluster elements and only them must all meet the given
Clustering approaches

Hierachical and pyramidal (conceptual) clustering
Non-Hierachical clustering

Generality Measure

Conceptual clustering method: Generality degree

Numerical criterion: measures the **Generality** of a description

- Clusters associated with less general descriptions should be formed first
- Set-valued and Interval-valued variables: evaluates the proportion of the description space covered
- Distribution-valued variables: evaluates the affinity between the given distribution and the Uniform
- Computed variable-wise; values combined in a multiplicative way give a measure of the variability of the description
- Extended to constrained data (rules between variables) with De Carvalho (1999, 2002)
Conceptual clustering method: Generality degree

Interval-valued variables: \( G(d_i) = \prod_{j=1}^{p} \frac{(\overline{l}_{ij} - l_{ij})}{L(O_j)} \)

\( L(O_j) \): total length of \( O_j \)

Set-valued variables: \( G(d_i) = \prod_{j=1}^{p} \frac{\#V_{ij}}{\#O_j} \)

Distribution-valued variables: evaluates the affinity between the given distribution and the Uniform

\( p_{ij\ell} \): weight of category \( \ell \) of variable \( j \) for entity \( i \)

Generalization by the maximum:

\( G_1(d_i) := \prod_{j=1}^{p} \frac{1}{\sqrt{k_j}} \sum_{\ell=1}^{k_j} \sqrt{p_{ij\ell}} \)

Generalization by the minimum:

\( G_2(d_i) := \prod_{j=1}^{p} \frac{1}{\sqrt{k_j(k_j-1)}} \sum_{\ell=1}^{k_j} \sqrt{1 - p_{ij\ell}} \)
Conceptual clustering: the algorithm

- Starting with the one-object clusters \( \{s_i\}, i = 1, \ldots, n \)
- At each step, form a cluster \( C \) union of \( C_1, C_2 \)
- \( C \) represented by \( d \)
  - Such that:
    - \( C_1, C_2 \) can be merged together
    - \( d \) is more general than \( d_1, d_2 \) (obtained by Generalization)
    - \( \text{Int}(C) = d \) and \( \text{Ext}_S(d) = C : (C, d) \) is a concept
    - \( G(d) \) is minimum
Conceptual clustering: recent approach

- Polaillon & Brito (2011): common framework for numerical (real or interval-valued), ordinal and distribution-valued variables
  - Generalization operator determines intents by intervals of values
- Variables of different types be taken together into account
- Distribution data: concepts more homogeneous than those obtained with the maximum or minimum operators, e.g.
  \[(0, 15], [0, 5); [15, 30], [0, 5) \cup_{int} ([0, 15], 0, 2; [15, 30], 0, 8) = ([0, 15][([0, 2, 0, 5); [15, 30][([0, 5, 0, 8])]
- Approach applied to hierarchical (or pyramidal) clustering (Brito and Polaillon (2012))
- Updating the “generality degree” - now additive on the variables: average variability
Real and Interval-Valued variables

$Y_j : S \rightarrow I, Y_j(s_i) = [l_{ij}, u_{ij}]$ ; $I$: set of intervals of $\mathbb{R}$

Generalization by interval union:

Intent of a st $A$ :

$d = (I_1, \ldots, I_p), I_j = [\text{Min}\{l_{ij}\}, \text{Max}\{u_{ij}\}], s_i \in A \subseteq S$
Generalization by Intervals: example

Variables: Age, Salary during the 5 recent years

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>30</td>
<td>[1000, 3000]</td>
</tr>
<tr>
<td>$s_2$</td>
<td>37</td>
<td>[1200, 3500]</td>
</tr>
<tr>
<td>$s_3$</td>
<td>28</td>
<td>[1500, 4000]</td>
</tr>
<tr>
<td>$s_4$</td>
<td>40</td>
<td>[2000, 3200]</td>
</tr>
</tbody>
</table>

$A = \{s_1, s_2, s_3\}$

Intent: $d = ([28, 37], [1000, 4000])$

Extent = $\{s_1, s_2, s_3\}$

$\implies C = (\{s_1, s_2, s_3\}, ([28, 37], [1000, 4000]))$ is a concept
Generalization by Intervals

Distributional variables

\( Y_1, \ldots, Y_p: p \) distributional variables

\( O_j = \{ c_{j1}, \ldots, c_{jk_j} \} \) set of \( k_j \) possible categories or sub-intervals of variable \( Y_j \)

\( M_j: \) set of distributions on \( O_j; M = M_1 \times \ldots \times M_p \)

\( Y_j(s_i) = \{ c_{j1}(p_{j1}^{s_i}), \ldots, c_{jk_j}(p_{jk_j}^{s_i}) \} \)

\( p_{jk_\ell}^{s_i}: \) probability/frequency associated with \( c_{j \ell} \) of \( Y_j \) and \( s_i \)
Generalization by Intervals

\[ A = \{s_1, \ldots, s_h\} \subseteq S \]

Intent:
\[ d_j = \{c_{j1}(l_{j1}), \ldots, c_{jk_j}(l_{jk_j})\} \]

\[ l_{j\ell} = \left[ \text{Min}\{p_{j\ell}^{s_i}\}, \text{Max}\{p_{j\ell}^{s_i}\} \right], s_i \in A \]

Extent:
\[ \{s_i \in S : p_{j\ell}^{s_i} \in l_{j\ell}\} \]
Generalization by Intervals: example

Categorical modal variables
Groups of students for each of which a categorical mark is given:
- a: mark < 10
- b: mark between 10 and 15
- c: mark > 15

<table>
<thead>
<tr>
<th>Group</th>
<th>Mark</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>&lt; 10(0.2), [10 – 15] (0.6), &gt; 15(0.2)</td>
</tr>
<tr>
<td>G2</td>
<td>&lt; 10(0.3), [10 – 15] (0.3), &gt; 15(0.4)</td>
</tr>
<tr>
<td>G3</td>
<td>&lt; 10(0.1), [10 – 15] (0.6), &gt; 15(0.3)</td>
</tr>
<tr>
<td>G4</td>
<td>&lt; 10(0.3), [10 – 15] (0.6), &gt; 15(0.1)</td>
</tr>
</tbody>
</table>

Generalization by intervals of $A = \{G1,G2\}$ provides the intent
Intent : $d = \{a \ [0.2, 0.3] , b \ [0.3, 0.6] , c \ [0.2, 0.4]\}$

The extent is $\{G1,G2\}$

$C = (\{G1, G2\}, (a \ [0.2, 0.3] , b \ [0.3, 0.6] , c \ [0.2, 0.4]))$ is a concept
Generalization by Intervals: example

**Ordinal variables**
Four cinema critics evaluate three movies:

<table>
<thead>
<tr>
<th></th>
<th>Movie 1</th>
<th>Movie 2</th>
<th>Movie 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critic 1</td>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Critic 2</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Critic 3</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Critic 4</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Intent of (Critic1, Critic2) = ([5, 5], [4, 5], [4, 4])
Extent = {Critic1, Critic2}

Intent of (Critic3, Critic4) = ([1, 2], [1, 2], [1, 2])
Extent = {Critic3, Critic4}
Generalization by Intervals: mixed example

Persons described by
Age - real-valued variable
Time (in minutes) they take to go to work - interval-valued variable
Means of transportation used - categorical modal variable
Classifications given to three newspapers, A, B and C - ordinal variables

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Time</th>
<th>Transport</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albert</td>
<td>25</td>
<td>[15, 20]</td>
<td>car (0.2) bus (0.8))</td>
<td>4</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Bellinda</td>
<td>40</td>
<td>[25, 30]</td>
<td>car (0.7), bus (0.2), train (0.1))</td>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Christine</td>
<td>32</td>
<td>[10, 15]</td>
<td>car (0.2), bus (0.7), train (0.1))</td>
<td>5</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>David</td>
<td>58</td>
<td>[30, 45]</td>
<td>car (0.9), bus (0.1))</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

**Intent** of \( A = \{\text{Albert, Christine}\} \) is

\( V = ([25, 32] , [10, 20] , ([0.2, 0.2] , [0.7, 0.8] , [0.0, 0.1]) , [4, 5] , [1, 2] , [4, 5]) \)

\((A, V)\) is a **concept**
Measuring Generality

Previously:

- Set-valued variables:
  proportion of the description space covered by d

- Distributional variables:
  affinity between the given distribution and the Uniform (Brito and De Carvalho (2008))

Now:

Measuring generality of a description $d$, $G(d)$ in a common manner for numerical (real and interval-valued), ordinal and distributional variables
Generality of a description $d = (d_1, \ldots, d_p)$ is evaluated variable by variable.

For variable $Y_j$ a value $G(d_j) \in [0, 1]$ is computed - measures proportion of description space $O_j$ covered by $d_j$.

The generality of a description is the arithmetic mean of the variable-wise values:

$$G(d) = \frac{1}{p} \sum_{j=1}^{p} G(d_j)$$
Measuring Generality

- $G(d_j)$ depends on the type of variable:
  - measure of the set covered by $d_j$
  - increasing as relates to inclusion

- Numerical variables: $Y_j : S \rightarrow [L, U] \quad d_j = [l_j, u_j]$
  
  
  $G(d_j) = \frac{u_j - l_j}{U - L}$

- Analogously for ordinal variables

- Distributional variables: $Y_j : S \rightarrow M_j$
  
  $d_j = \{c_{j1}(l_{j1}), \ldots, c_{jk}(l_{jk})\}, \quad l_{j\ell} = [l_{j\ell}, l_{j\ell}]$

  $G(d_j) = \frac{1}{k_j} \sum_{\ell=1}^{k_j} (l_{j\ell} - l_{j\ell})$
Measuring Generality: Example

Two groups $G_1, G_2$ described by
$Y_1$: age group, categorical modal variable $Y_2$: salary, interval-valued variable, $Y_2: S \rightarrow [0, 10]$

$G_1: (a(0.2), b(0.6), c(0.2), [2, 5])$

$G_2: (a(0.3), b(0.3), c(0.4)), [1, 2.5])$

The joint description of the 2 groups is :
$d = (a [0.2, 0.3], b [0.3, 0.6], c [0.2, 0.4], [1, 5])$

$G(d_1) = \frac{1}{3} ((0.3 - 0.2) + (0.6 - 0.3) + (0.4 - 0.2)) = 0.2$

$G(d_2) = \frac{5-1}{10-0} = 0.4$

$G(d) = \frac{1}{2} (0.2 + 0.4) = 0.3$
Conceptual clustering: Application

Age distribution and salary range of several groups

Five groups described by:

\[ Y_1 = \text{Age class}, \ a: \ \text{age} < 25, \ b: \ \text{age} \in [25, 60], \ c: \ \text{age} > 60 \]
\[ Y_2 = \text{Salary}, \ Y_2 : E \rightarrow [0, 10] \]

<table>
<thead>
<tr>
<th>Group</th>
<th>Age</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>( G_1 )</td>
<td>( a(0.2), b(0.6), c(0.2) )</td>
<td>( [2, 5] )</td>
</tr>
<tr>
<td>( G_2 )</td>
<td>( a(0.3), b(0.3), c(0.4) )</td>
<td>( [1, 2.5] )</td>
</tr>
<tr>
<td>( G_3 )</td>
<td>( a(0.1), b(0.6), c(0.3) )</td>
<td>( [3, 6] )</td>
</tr>
<tr>
<td>( G_4 )</td>
<td>( a(0.3), b(0.6), c(0.1) )</td>
<td>( [4, 8] )</td>
</tr>
<tr>
<td>( G_5 )</td>
<td>( a(0.5), b(0.3), c(0.2) )</td>
<td>( [1.5, 3] )</td>
</tr>
</tbody>
</table>
Application: the conceptual indexed hierarchy

The concepts are:

1. \(C^{(6)} = \{G2, G5\} \); \(d(6) = \{a([0.3, 0.5]), b([0.3, 0.3]), c([0.2, 0.4])\}, [1, 3]\); \(G(d^{(6)}) = 0.17\)

2. \(C^{(7)} = \{G1, G3\} \); \(d(7) = \{a([0.1, 0.2]), b([0.6, 0.6]), c([0.2, 0.3])\}, [2, 6]\); \(G(d^{(7)}) = 0.23\)

3. \(C^{(8)} = \{G1, G3, G4\} \); \(d(8) = \{a([0.1, 0.2]), b([0.5, 0.6]), c([0.1, 0.3])\}, [2, 8]\); \(G(d^{(8)}) = 0.4\)

4. \(C^{(9)} = \{G1, G2, G3, G4, G5\} \); \(d(9) = \{a([0.1, 0.5]), b([0.3, 0.6]), c([0.1, 0.4])\}, [1, 8]\); \(G(d^{(9)}) = 0.5\)
Clustering approaches

Divisive Clustering

The criterion
The distances
Binary questions and Assignment
The algorithm

Hierachical and pyramidal (conceptual) clustering

Generality Measure

Non-Hierachical clustering
SCLUST: Dynamical clustering for symbolic data (De Carvalho et al. (2008))

SCLUST: non-hierarchical clustering on symbolic data, using a k-means - or dynamical clustering - like method

- Starting from a partition on a pre-fixed number of clusters
- alternates an assignment step (based on minimum distance to cluster prototypes)
- and a representation step (which determines new prototypes in each cluster)
- until convergence is achieved (or a pre-fixed number of iterations is reached)
SCLUST: Dynamical clustering for symbolic data

Define $D(A, c) = \sum_{s \in A} d(s, c)$

Assigning function $f(c_1, \ldots, c_k) = \{P_1, \ldots, P_k\}$

Given the centers $(c_1, \ldots, c_k)$
the partition $P = \{P_1, \ldots, P_k\}$ is defined by:

$P_h = \{s \in S : D(\{s\}, c_h) \leq D(\{s\}, c_m), 1 \leq m \leq k\}$

Representation function $g\{P_1, \ldots, P_k\} = (c_1, \ldots, c_k)$

Given a partition $\{P_1, \ldots, P_k\}$,
the centers $(c_1, \ldots, c_k)$ are defined by:
$c_h : D(P_h, c_h)$ minimizes $D(P_h, \bullet)$
SCLUST: Dynamical clustering for symbolic data

In each step:

Decrease of the value of a criterion that evaluates the distance of each element $s_i$ to the center of its class $P_\ell$

$\rightarrow$ evaluates the fit between the classes $P_\ell$ and their representants $c_\ell$

$$W = \sum_{\ell=1}^{k} \sum_{s \in P_\ell} d(s, c_\ell)$$

From the consistency between functions $f$ and $g$, and the assurance of the existence and unicity of the centers determined by $g$, it follows that $W$ decreases in each step, converging to a local optimum.
SCLUST: Dynamical clustering for symbolic data

- The method locally optimizes a criterion that measures the fit between cluster prototypes and cluster members, which is additive, and based on the assignment-distance function.

- The method allows for all types of variables in the input data.

- Selects the distances for the assigning step accordingly:
  - Quantitative real-valued data: Euclidean distance
  - Interval and quantitative multi-valued data: Hausdorff distance
  - Categorical single-valued data: $\chi^2$-square distance
  - Categorical multi-valued data: De Carvalho distance
  - Distributional data: a classical $\phi^2$ distance

SCLUST includes functions for the determination of the appropriate number of clusters, based on classical indices (see Hardy, (2008)).