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. Hierachical and pyramidal (conceptual) clustering
   - Generality Measure
4. Non-Hierachical clustering

Outline

1. Clustering approaches
2. Divisive Clustering
   - The criterion
   - The distances
   - Binary questions and Assignement
   - The algorithm
3. Hierachical and pyramidal (conceptual) clustering
   - Generality Measure
4. Non-Hierachical 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 litterature
- Interval-valued data:
  - Minkowski-type distances
  - Malahanobis distance
  - Hausdorff distance
- Distribution-valued data
  - Wasserstein distance
  - Mallows distance

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 & Lechevalllier (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:
  - Bock *et al* (2002)
  - De Carvalho *et al* (2011)
  - Hajjar and Hamdan (2011)
  - Yang, Hung, Chen (2012)

Interval-valued variables: Distance measures

Many measures proposed in the litterature

Hausdorff distance :

$$d_H(I_i, I_j) = \max\left\{\left\{|l_i - l_j|\,, |u_i - u_j|\right\}\right.$$

Euclidean distance :

$$d_2(I_i, I_j) = \sqrt{(l_i - l_j)^2 + (u_i - u_j)^2}$$

City-Block distance :

$$d_1(I_i, I_j) = |l_i - l_j| + |u_i - u_j|\,.$$

Malahanobis distance:

defined on the basis of the vectors of observed lower $X_{iL} = (l_{i1}, \ldots, l_{ip})$ and upper bounds $X_{iU} = (u_{i1}, \ldots, u_{ip})$.

$d(s_{i_1}, s_{i_2}) = d_M(X_{i_1 L}, X_{i_2 L}) + d_M(X_{i_1 U}, X_{i_2 U})$ where

$d_M(X_{i_1 L}, X_{i_2 L}) = (X_{i_1 L} - X_{i_2 L})^t M_L (X_{i_1 L} - X_{i_2 L})$ is the Mahalanobis distance between the two vectors $X_{i_1 L}$ and $X_{i_2 L}$

$d_M(X_{i_1 U}, X_{i_2 U}) = (X_{i_1 U} - X_{i_2 U})^t M_U (X_{i_1 U} - X_{i_2 U})$

Clustering distributional data

- Hardy (2004, 2008) developped SHICLUST
    - extends single and complete linkage, centroid and Ward methods to categorical modal variables
    - dissimilarity measures and aggregation indices adapted or suitably chosen
- Verde and Irpino (2006, 2007, 2008)
    - 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

Histogram-valued variables: Distance measures

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

Divergency measures	
Kullback-Leibler	$D_{KL}(f,g) = \int_{\mathbb{R}} log\left(\frac{f(x)}{g(x)}\right) f(x)dx$
Jeffrey	$D_J(f,g) = D_{KL}(f,g) + D_{KL}(g,f)$
χ^2	$D_{\chi^2}(f,g) = \int_{\mathbb{R}} \frac{\lvert f(x)-g(x)\rvert^2}{g(x)} dx$
Hellinger	$D_H(f,g) = \left[\int_{\mathbb{R}} \left(\sqrt{f(x)} - \sqrt{g(x)}\right) dx\right]^{\frac{1}{2}}$
Total variation	$D_{var}(f,g) = \int_{\mathbb{R}} \lvert f(x)-g(x)\rvert dx$
Wasserstein	$D_W(f,g) = \int_{\mathbb{R}} \lvert F^{-1}(x) - G^{-1}(x)\rvert dx$
Mallows	$D_M(f,g) = \sqrt{\int_0^1 \left(F^{-1}(x) - G^{-1}(x)\right)^2 dx}$
Kolmogorov	$D_W(f,g) = \max_{\mathbb{R}} \lvert F(x) - G(x)\rvert$

Outline

1. Clustering approaches
2. Divisive Clustering
   - The criterion
   - The distances
   - Binary questions and Assignement
   - The algorithm
3. Hierachical and pyramidal (conceptual) clustering
   - Generality Measure
4. Non-Hierachical 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 $P_m = \left\{ C_1^{(m)}, C_2^{(m)}, \ldots, C_m^{(m)} \right\}$ 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:

$$d_{H_2}(s_{i_1}, s_{i_2}) = \sqrt{\sum_{j=1}^{p} (\max\{|l_{i_1 j} - l_{i_2 j}|, |u_{i_1 j} - u_{i_2 j}|\})^2}$$

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 $p_{kj} = \dfrac{f_{kj}}{np}$

χ^2distance :

$$d(s_k, s_\ell) = \sum_{j=1}^{t} \frac{p_{..}}{p_{.j}} \left(\frac{p_{kj}}{p_{k.}} - \frac{p_{\ell j}}{p_{\ell .}} \right)^2$$

Distance measures: New approach

Evaluate the dissimilarity between distributions

$$Y_j(S_i) = H_{Y_j(S_i)} = ([\underline{I}_{ij1}, \overline{I}_{ij1}[, p_{ij1}; \ldots ; [\underline{I}_{ijK_j}, \overline{I}_{ijK_j}], p_{ijK_j})$$

1. **Mallows distance**

$$d^2_M(x_{ij}, x_{i'j}) = \int_0^1 (q_{ij}(t) - q_{i'j}(t))^2 dt$$

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

2. **Squared Euclidean distance**

$$d^2_E(x_{ij}, x_{i'j}) = \sum_{\ell=1}^{K_j} (p_{ij\ell} - p_{i'j\ell})^2$$

Binary questions

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

- Numerical data :
  $R_{j\ell} := Y_j \in R_{1j} \Leftrightarrow Y_j \leq m_\ell, j = 1, \ldots, p$
- 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 Assignement

- Interval variables in SODAS:
  - m_ℓ - 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_ℓ - defining the "cuts" are the $\overline{I}_{j\ell}$
  - $R_{j\ell} : Y_j \leq \overline{I}_{j\ell}$ 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 assignement: example

	Age	Marks
Class 1	$([10, 11[, 0.50;$ $[11, 12[, 0.50;$ $[12, 14], 0.00)$	$([5, 10[, 0.2; [10, 12[, 0.5; [12, 14[, 0.133;$ $[14, 15[, 0.067; [15, 16[, 0.033;$ $[16, 18[, 0.067; [18, 19], 0.0)$
Class 2	$([10, 11[, 0.00;$ $[11, 12[, 0.33;$ $[12, 14], 0.67)$	$([5, 10[, 0.05; [10, 12[, 0.3; [12, 14[, 0.25;$ $[14, 15[, 0.1; [15, 16[, 0.1;$ $[16, 18[, 0.133; [18, 19], 0.067)$

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

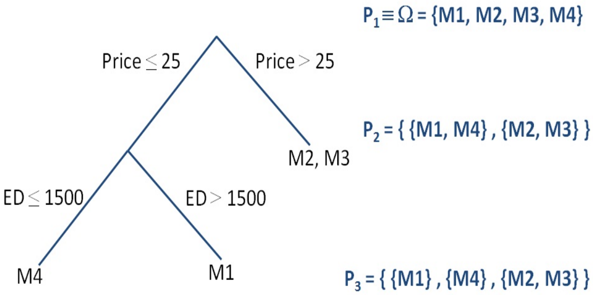
Divisive Clustering: Application

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

	Price	Engine Displacement
Model 1	$([15, 25[, 0.5; [25, 35[, 0.5)$;	$([1300, 1500[, 0.2; [1500, 1700[, 0.5;$ $[1700, 1900[, 0.3)$
Model 2	$([15, 25[, 0.2; [25, 35[, 0.8)$;	$([1300, 1500[, 0.1; [1500, 1700[, 0.2;$ $[1700, 1900[, 0.7)$
Model 3	$([15, 25[, 0, 33; [25, 35[, 0.67)$	$([1300, 1500[, 0.1; [1500, 1700[, 0.4;$ $[1700, 1900[, 0.5)$
Model 4	$([15, 25[, 0.6; [25, 35[, 0.4)$	$([1300, 1500[, 0.6; [1500, 1700[, 0.4;$ $[1700, 1900[, 0.0)$

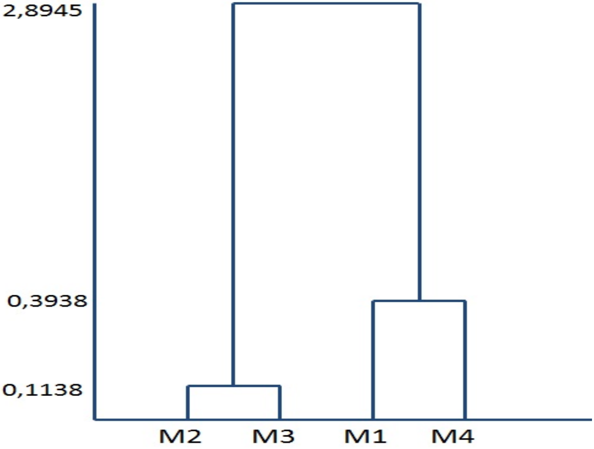
- Partition into three clusters
- Squared Euclidean distance between distributions to compare the observed values for each car model

Divisive clustering application: the clustering tree


P₁ ≡ Ω = {M1, M2, M3, M4}
Price ≤ 25
Price > 25
P₂ = { {M1, M4} , {M2, M3} }
M2, M3
ED ≤ 1500
ED > 1500
M4
M1
P₃ = { {M1} , {M4} , {M2, M3} }

- Cluster $C_1^{(3)} = \{\text{Model } 4\}$: "Price $\leq 25 \wedge$ Engine Displacement ≤ 1500"
- Cluster $C_2^{(3)} = \{\text{Model } 1\}$: "Price $\leq 25 \wedge$ Engine Displacement > 1500"
- Cluster $C_3^{(3)} = \{\text{Model } 2, \text{ Model } 3\}$: "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

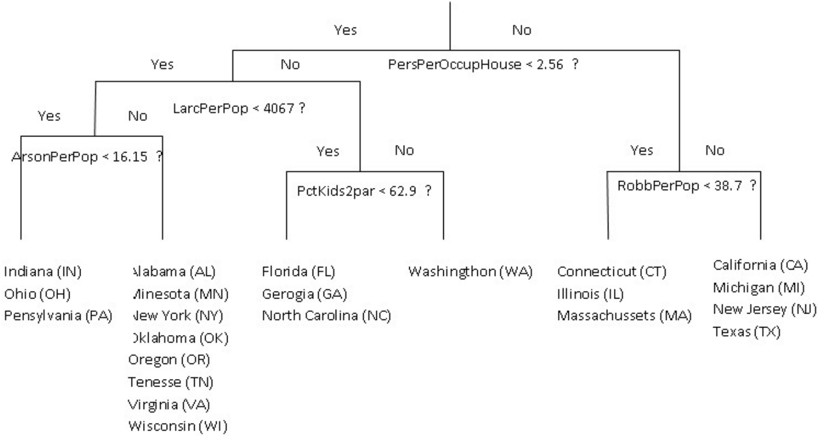
CRIMES

- murdPerPop: number of murders per 100K population
- robbbPerPop: 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 popuation

Crime-data application: the dendrogram



Outline

1. Clustering approaches
2. Divisive Clustering
   - The criterion
   - The distances
   - Binary questions and Assignement
   - The algorithm
3. Hierachical and pyramidal (conceptual) clustering
   - Generality Measure
4. Non-Hierachical clustering

Hierachical/PyramidalConceptual clustering method (Brito (1991, 1995))

- Ascending hierarchical / pyramidal clutering
- 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. $([0, 15[, 0, 5; [15, 30[, 0, 5) \cup_{max} ([0, 15[, 0, 2; [15, 30[, 0, 8) = ([0, 15[, 0, 5; [15, 30[, 0, 8)$
- Only clusters corresponding to concepts are formed : the cluster elements and only them must all meet the given

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: $G(d_i) = \prod_{j=1}^{p} \frac{(\overline{I_{ij}} - \underline{I_{ij}})}{L(O_j)}$
$L(O_j)$: 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 ℓ 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)
  - $Int(C) = d$ and $Ext_S(d) = C$: (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))
- Updapting the "generality degree" - now additive on the variables: average variability

Generalization by Intervals

Real and Interval-Valued variables

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

Generalization by interval union:

Intent of a st A :

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

Generalization by Intervals: example

Variables : Age, Salary during the 5 recent years

	Age	Salary
s_1	30	$[1000, 3000]$
s_2	37	$[1200, 3500]$
s_3	28	$[1500, 4000]$
s_4	40	$[2000, 3200]$

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

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

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

$$\Longrightarrow C = (\{s_1, s_2, s_3\}, ([28, 37], [1000, 4000]) \text{ is a concept}$$

Generalization by Intervals

Distributional variables

$Y_1, \dots, Y_p$: p distributional variables

$O_j = \left\{ c_{j1}, \dots, c_{jk_j} \right\}$ set of k_j possible categories or sub-intervals of variable Y_j

M_j : set of distributions on O_j ; $M = M_1 \times \dots \times M_p$

$Y_j(s_i) = \left\{ c_{j1}(p_{j1}^{s_i}), \dots, c_{jk_j}(p_{jk_j}^{s_i}) \right\}$

$p_{jk_\ell}^{s_i}$: 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 :

$d_j = \{c_{j1}(I_{j1}), \ldots, c_{jk_j}(I_{jk_j})\}$

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

Extent :

$\{s_i \in S : p_{j\ell}^{s_i} \in I_{j\ell}\}$

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 :

	Mark
Group 1 (G1)	$< 10(0.2), [10-15]\,(0.6), > 15(0.2)$
Group 2 (G2)	$< 10(0.3), [10-15]\,(0.3), > 15(0.4)$
Group 3 (G3)	$< 10(0.1), [10-15]\,(0.6), > 15(0.3)$
Group 4 (G4)	$< 10(0.3), [10-15]\,(0.6), > 15(0.1)$

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:

	Movie 1	Movie 2	Movie 3
Critic 1	5	5	4
Critic 2	5	4	4
Critic 3	1	2	2
Critic 4	2	1	1

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

	Age	Time	Transport	A	B	C
Albert	25	$[15, 20]$	car (0.2) bus (0.8))	4	2	5
Bellinda	40	$[25, 30]$	car (0.7), bus (0.2), train (0.1))	2	4	3
Christine	32	$[10, 15]$	car (0.2), bus (0.7), train (0.1))	5	1	4
David	58	$[30, 45]$	car (0.9), bus (0.1))	2	4	1

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

Measuring Generality

- 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 couvered 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 : $Y_j : S \to [L, U] \; d_j = [l_j, u_j]$

  $$G(d_j) = \frac{u_j - l_j}{U - L}$$
- Analogously for ordinal variables
- Distributional variables : $Y_j : S \to M_j$

  $$d_j = \{c_{j1}(I_{j1}), \ldots, c_{jk_j}(I_{jk_j})\}, \; I_{j\ell} = \left[\underline{I_{j\ell}}, \overline{I_{j\ell}}\right]$$

  $$G(d_j) = \frac{1}{k_j} \sum_{\ell=1}^{k_j} (\overline{I_{j\ell}} - \underline{I_{j\ell}})$$

Measuring Generality: Example

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

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

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

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

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

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

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

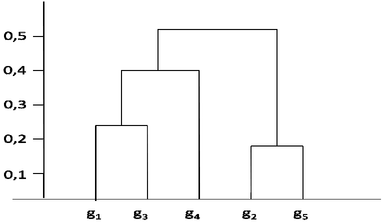
Conceptual clustering: Application

Age distribution and salary range of several groups

Five groups described by:
Y_1 = Age class, a: age < 25 , b: age $\in [25, 60]$, c: age > 60
Y_2 = Salary, $Y_2 : E \rightarrow [0, 10]$

Group	Age	Salary
G_1	$a(0.2), b(0.6), c(0.2)$	$[2, 5]$
G_2	$a(0.3), b(0.3), c(0.4)$	$[1, 2.5]$
G_3	$a(0.1), b(0.6), c(0.3)$	$[3, 6]$
G_4	$a(0.3), b(0.6), c(0.1)$	$[4, 8]$
G_5	$a(0.5), b(0.3), c(0.2)$	$[1.5, 3]$

Application: the conceptual indexed hierarchy



The concepts are :

- $C^{(6)} = \{G2, G5\}$; $d(6) = (\{a([0.3, 0.5]), b([0.3, 0.3]), c([0.2, 0.4])\}, [1, 3])$; $G(d^{(6)}) = 0.17$
- $C^{(7)} = \{G1, G3\}$; $d(7) = (\{a([0.1, 0.2]), b([0.6, 0.6]), c([0.2, 0.3])\}, [2, 6])$; $G(d^{(7)}) = 0.23$
- $C^{(8)} = \{G1, G3, G4\}$; $d(8) = (a([0.1, 0.2]), b([0.5, 0.6]), c([0.1, 0.3]), [2, 8])$; $G(d^{(8)}) = 0.4$
- $C^{(9)} = \{G1, G2, G3, G4, G5\}$;
  $d(9) = (\{a([0.1, 0.5]), b([0.3, 0.6]), c([0.1, 0.4])\}, [1, 8])$; $G(d^{(9)}) = 0.5$

Outline

1. Clustering approaches
2. Divisive Clustering
   - The criterion
   - The distances
   - Binary questions and Assignement
   - The algorithm
3. Hierachical and pyramidal (conceptual) clustering
   - Generality Measure
4. Non-Hierachical clustering

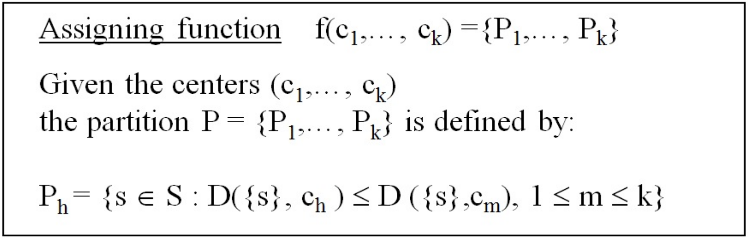
SCLUST: Dynamical clustering for symbolic data (De Carvalho et al. (2008))

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 protoypes in each cluster)
- until convergence is achieved (or a pre-fixed number of iterations is reached)

SCLUST: Dynamical clustering for symbolic data

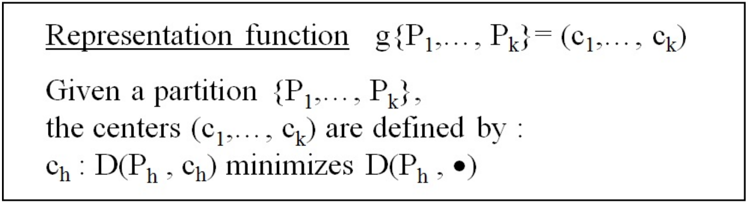
Define $D(A, c) = \sum_{s \in A} d(s, c)$



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_ℓ

$\longrightarrow$ evaluates the fit between the classes P_ℓ and their representants c_ℓ

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

From the consistency between functions f and g, and the assurance of the existance 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 propotypes 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: χ-square distance
  - Categorical multi-valued data: De Carvalho distance
  - Distributional data: a classical ϕ^2 distance

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