Symbolic Clustering Based on Quantile Representation

Paula Brito
Universidade do Porto
Portugal

Manabu Ichino
Tokyo Denki University
Japan

Outline

- Objective
- Symbolic variables
- The m-quantile representation model
  - Interval-valued variables
  - Histogram-valued variables
  - Categorical multi-valued variables
- Conceptual clustering based on the quantile representation
  - The criterion
  - The aggregation by mixture
  - Representation of the new cluster
- Illustrative example

Quantile representation

Objective:

- Obtain a common representation model for different variable types

- Allowing to apply clustering methods to the full (originally) mixed data array

Symbolic Variables

- Symbolic data → new variable types:
  - Set-valued variables : variable values are subsets of an underlying set
    - Interval variables
    - Categorical multi-valued variables
  - Modal variables : variable values are distributions on an underlying set
    - Histogram variables

Symbolic Variables

Let $Y_1, \ldots, Y_p$ be the variables,

O_j the underlying domain of Y_j

B_j the observation space of Y_j , $j=1, \ldots, p$: $Y_j : \Omega \rightarrow B_j$

- Y_j classical variable : $B_j = O_j$
- Y_j interval variable : B_j set of intervals of O_j
- Y_j categorical multi-valued variable : $B_j = P(O_j)$
- Y_j modal variable : B_j set of distributions on O_j

Symbolic data array

The dataset consists of information's about patients (adults) in healthcare centers, during the second semester of 2008.

Healthcare Center	Sex Y_2	Age Y_3	Degree Y_4	Emergency consults Y_5	Waiting time for consult (in minutes) Y_6	Pulse Y_7
A	{F, 1; M, 3}	[25,53]	{9th grade, 1/2; Higher education, 1/2}	{0,1,2}	{[0,15[, 0;[15,30[, 0.25;[30,45[, 0.5; [45,60[,0;≥60,0.25}	[44,86]
B	{F, 3; M, 1}	[33,68]	{6th grade, 1/4; 9th grade, 1/4; 12th grade, 1/4; ; Higher education,, 1/4}	{1,4,5,10}	{[0,15[, 0.25; [15,30[, 0.25; [30,45[, 0.25; [45,60[,0.25;≥60,0}	[54,76]
C	{F, 1; M, 2}	[20;75]	{4th grade, 1/3; 9th grade, 1/3; 12th grade 1/3}	{0,5,7}	{[0,15[, 0.33; [15,30[, 0;[30,45[, 0.33; [45,60[,0; ≥60,0.33}	[70,86]

Common representation model

- Use the m-quantiles of the underlying distribution of the observed data values (Ichino, 2008)

$$(\min\ ;\ Q_1;\ \ldots\ ;\ Q_{m-1}\ ;\ \max)$$

- When quartiles are chosen (m=4) , the representation for each variable is defined by the 5-uple

$$(\min\ ;\ Q_1\ ;\ Q_2\ ;\ Q_3\ ;\ \max)$$

$\Rightarrow$ Determination of quantiles for each variable type

Common representation model: determining quantiles

- Interval-valued variables

  An underlying distribution is assumed within each observed interval

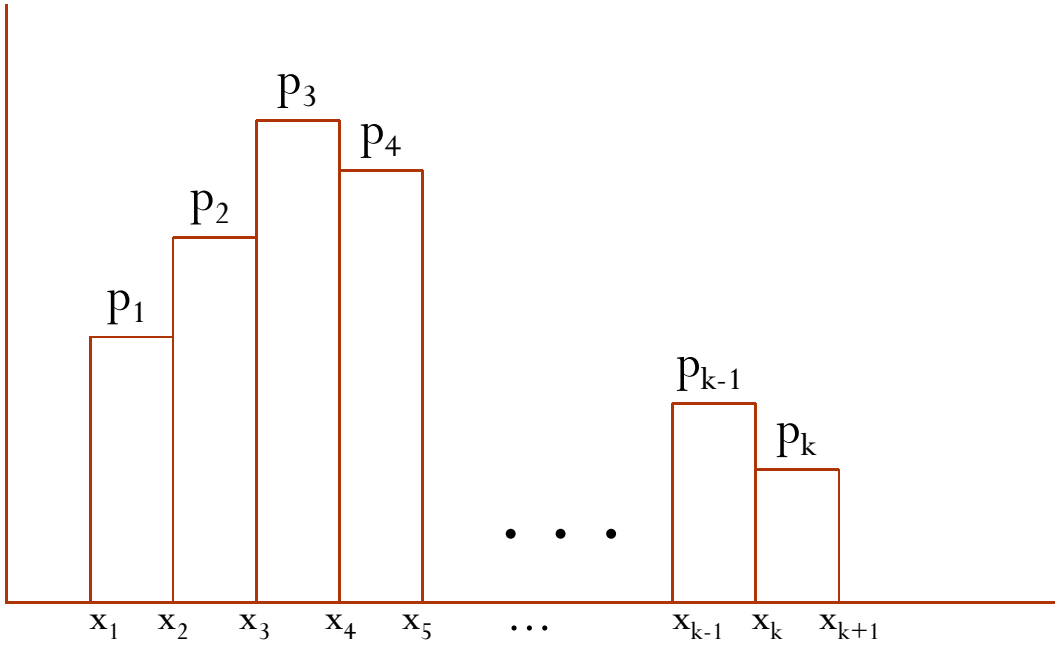
  - ✓ Uniform (Bertrand and Goupil, 2000)

$$Y_j(\omega_i) = [l_{ij}, u_{ij}]$$

$$Q_q = l_{ij} + \frac{q}{m}(u_{ij} - l_{ij}),\ q = 1, \ldots m-1$$

Common representation model: determining quantiles

- Histogram-valued variables
  - Quantiles obtained by interpolation
  - Uniform distribution assumed in each class (bid)



Common representation model: determining quantiles

- Categorical multi-valued variables

Y categorical multi-valued variable taking possible k categories

$c_\ell \quad \ell = 1, 2, \ldots, k$

p_ℓ - relative frequency of category c_ℓ for the n objects

Rank the categories $c_1, c_2, \ldots, c_k$ according to the frequency values p_ℓ .

Define a uniform cumulative distribution function for each object $\omega_i \in \Omega$ based on the ranking, assuming continuity.

Then find the m-1 quantile values.

Example: Oils data

Oils \ Variables	Specific gravity (g/cm3)	Freezing point (°C)	Iodine value	Saponification value	Major acids
Linseed	[0.930 , 0.935]	[-27 , -18]	[170 , 204]	[118 , 196]	L, Ln, O, P, M
Perilla	[0.930 , 0.937]	[-5 , -4]	[192 , 208]	[188 , 197]	L, Ln, O, P, S
Cotton	[0.916 , 0.918]	[-6 , -1]	[99 , 113]	[189 , 198]	L, O, P, M,S
Sesame	[0.920 , 0.926]	[-6 , -4]	[104 , 116]	[187 , 193]	L, O, P, S, A
Camelia	[0.916 , 0.917]	[-21 , -15]	[80 , 82]	[189 , 193]	L, O
Olive	[0.914 , 0.919]	[0, 6]	[79 , 90]	[187 , 196]	L, O, P, S
Beef	[0.860 , 0.870]	[30, 38]	[40 , 48]	[190 , 199]	O, P, M, C,S
Hog	[0.858 , 0.864]	[22, 32]	[53 , 77]	[190 , 202]	L, O, P, M, S, Lu

Oils data : Quartile representation

Oil \ Acid	Lu	A	C	Ln	M	S	P	L	O
Linseed	0	0	0	0.2	0.2	0	0.2	0.2	0.2

Linseed : [0,1[: 0 ; [1,2[: 0 ; [2,4[: 0 ; [4,5[: 0.2 ;

[5,6[: 0.4 ; [6,7[: 0.4 ; [7,8[: 0.6 ; [8,9[: 0.8 ; [9,10[: 1

Min = 4 $Q_1 = 5.25$ $Q_2 = 7.5$ $Q_3 = 8.75$ Max = 10

		Spec. Grav.	Freezing P.	Iodine	Saponific.	M. Acids
Linseed	Min	0,93000	-27	170	118	4
	Q_1	0.93125	-24.75	178.5	137.5	5.25
	Q_2	0.93250	-22.5	187	157	7.5
	Q_3	0.93375	-20.25	195.5	176.5	8.75
	Max	0.93500	-18	204	196	10

Clustering methodology

- Standardization : $u'_{ij} = \frac{u_{ij} - Min_j}{Max_j - Min_j}$
- Data units compared by the Euclidean distance on the quantile vector representation
- Clusters also represented by a quantile vector
- Clusters also compared by the Euclidean distance

The algorithm

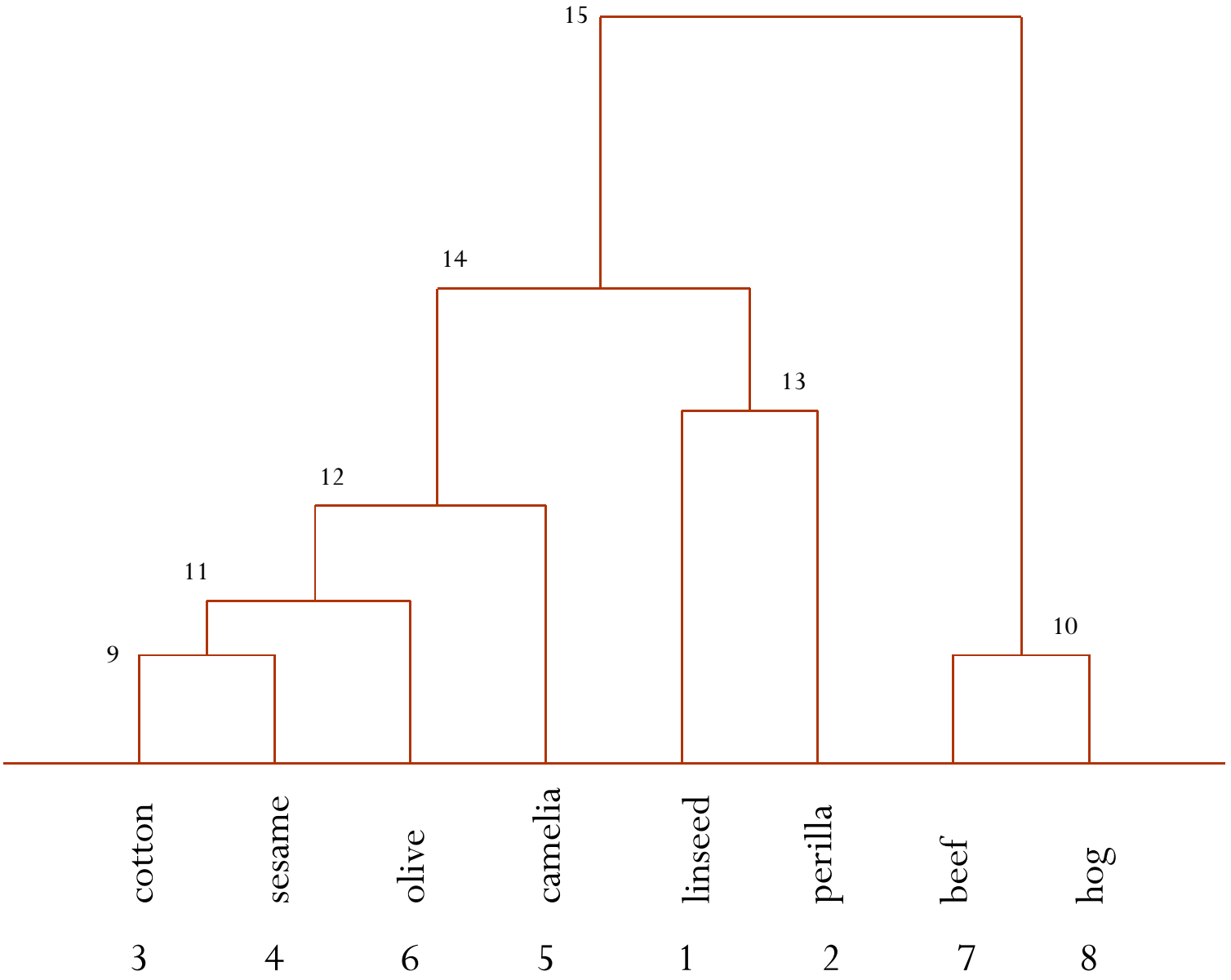
- Initial clusters are the single elements, each represented by a (m+1) quantile vector (min ; Q_1; … ; Q_{m-1} ; max)
- Choose the two clusters A and B with lowest Euclidean distance to be merged
- Assuming piecewise linear distributions, determine the distribution values of the quantiles of A on the distribution of the B, and vice-versa
- Take the mean of these distribution values on each of the $2 \times (m+1)$ points
- Assuming again piecewise linearity, determine the (m+1) quantiles of the new distribution, which represent the new cluster
- Iterate until a full hierarchy is obtained

Example: Oils data

Oils \ Variables	Specific gravity (g/cm3)	Freezing point (°C)	Iodine value	Saponification value	Major acids
Linseed	[0.930 , 0.935]	[-27 , -18]	[170 , 204]	[118 , 196]	L, Ln, O, P, M
Perilla	[0.930 , 0.937]	[-5 , -4]	[192 , 208]	[188 , 197]	L, Ln, O, P, S
Cotton	[0.916 , 0.918]	[-6 , -1]	[99 , 113]	[189 , 198]	L, O, P, M,S
Sesame	[0.920 , 0.926]	[-6 , -4]	[104 , 116]	[187 , 193]	L, O, P, S, A
Camelia	[0.916 , 0.917]	[-21 , -15]	[80 , 82]	[189 , 193]	L, O
Olive	[0.914 , 0.919]	[0, 6]	[79 , 90]	[187 , 196]	L, O, P, S
Beef	[0.860 , 0.870]	[30, 38]	[40 , 48]	[190 , 199]	O, P, M, C,S
Hog	[0.858 , 0.864]	[22, 32]	[53 , 77]	[190 , 202]	L, O, P, M, S, Lu

- Determination of quartile (m=4) representation for each variable
- Standardization

Classification of the oils



Cluster representation

Class number: 14 class A: 12 class B: 13 distance= 1.75100638182962

Specific gravity : (0.7088608, 0.7424438, 0.8607595, 0.9483122, 1)
Range= 0.2911392 ; IQD= 0.2058684

Freezing point : (0, 0.1107692, 0.1762238, 0.3510490, 0.5076923)
Range= 0.5076923 ; IQD= 0.2402797

Iodine value : (0.2321429, 0.2485119, 0.4523810, 0.927619, 1)
Range= 0.7678571 ; IQD= 0.6791071

Saponification value : (0, 0.8236486, 0.8656454, 0.894332, 0.952381)
Range= 0.952381 ; IQD= 0.07068347

Major acids : (0.1111111, 0.6023402, 0.7929029, 0.8990216, 1)
Range= 0.8888889 ; IQD= 0.2966813

Final remarks

- Common representation model for symbolic variables of different kinds
- Allows for clustering based on the full data description
- Clustering based on quantiles' proximity
- Uniformity assumed for the initial data
- Mixture of the distribution functions defined by the quantiles – piecewise linear functions
- Each new cluster is represented by the quantile vector obtained from the mixture (non-uniformity for clusters !)

Common representation model: determining quantiles

- **Histogram-valued variables**
- Distribution function:

$F(x) = 0$ for $x \leq x_1$

$F(x) = p_1(x-x_1)/(x_2-x_1)$ for $x_1 \leq x \leq x_2$

$F(x) = F(x_2) + p_2(x-x_2)/(x_3-x_2)$ for $x_2 \leq x \leq x_3$

……

$F(x) = F(x_k) + p_k(x-x_k)/(x_{k+1}-x_k)$ for $x_k \leq x \leq x_{k+1}$

$F(x) = 1$ for $x_{k+1} \leq x$

- Then find m+1 numerical values, the m-quantile values $y_1, y_2, \ldots, y_m, y_{m+1}$:

$F(y_1) = 0$, (i.e. $y_1 = x_1$)

$F(y_2) = 1/m$, $F(y_3) = 2/m$, … , $F(y_m) = (m-1)/m$, and

$F(y_{m+1}) = 1$, (i.e. $y_{m+1} = x_{k+1}$).

Oils data : ranking “major acids”

Oil \ Acid	Lu	A	C	Ln	M	S	P	L	O
Linseed	0	0	0	0.2	0.2	0	0.2	0.2	0.2
Perilla	0	0	0	0.2	0	0.2	0.2	0.2	0.2
Cotton	0	0	0	0	0.2	0.2	0.2	0.2	0.2
Sesame	0	0.2	0	0	0	0.2	0.2	0.2	0.2
Camelia	0	0	0	0	0	0	0	0.5	0.5
Olive	0	0	0	0	0	0.25	0.25	0.25	0.25
Beef	0	0	0.2	0	0.2	0.2	0.2	0	0.2
Hog	0.167	0	0	0	0.167	0.167	0.167	0.167	0.167
$\Sigma\, q_{i\ell}$	0.167	0.2	0.2	0.4	0.767	1.217	1.417	1.717	1.917
Rank	1	2	2	4	5	6	7	8	9