

DIV

A divisive and symbolic clustering method

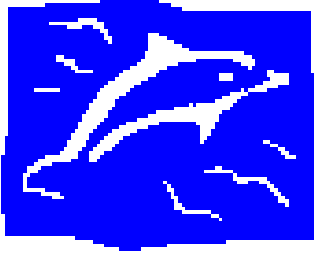
Marie CHAVENT

*M.A.B. Laboratory (UMR 5466)
University Bordeaux1
BORDEAUX*

Myriam TOUATI

*LISE-CEREMADE Laboratory (UMR 7534)
University PARIS IX Dauphine
PARIS*

FRANCE



Plan



- ⌘ DIV objectives
- ⌘ Input data
- ⌘ Method and algorithm
- ⌘ Output results
- ⌘ Example: data processing on ONS census data
- ⌘ Conclusion
- ⌘ References



DIV Objectives



⌘ Performs an hierarchy

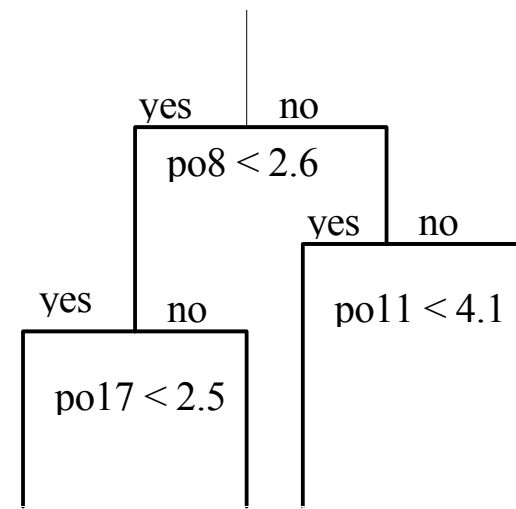
- ☒ From top to bottom
- ☒ By successive splitting of classes

⌘ Input:

- ☒ symbolic objects

⌘ Output:

- ☒ A hierarchy of partitions (2 to K clusters) = a decision tree
- ☒ Assertion object = clusters



Cluster 1 Cluster 2 Cluster 3 Cluster 4

Cluster 1 = $[po8 < 2.6] \wedge [po17 < 2.5]$

Cluster 2 = $[po8 < 2.6] \wedge [po17 > 2.5]$

Cluster 3 = $[po8 > 2.6] \wedge [po11 < 4.1]$

Cluster 4 = $[po8 > 2.6] \wedge [po11 > 4.1]$



Input data: classical or symbolic objects



EXAMPLE

⊞ Classical data matrix

⊞ continuous →

AGE = 23

⊞ ordinal →

HEIGHT = medium

⊞ Symbolic data matrix

⊞ interval →

AGE = [20 : 30]

⊞ multi-ordinal →

and HEIGHT = (small, large, very large)

⊞ binary →

and SEX = M

⊞ probabilistic →

and WEIGHT = (light(0.25), heavy (0.75))



Method: DIV algorithm



Divisive and symbolic algorithm

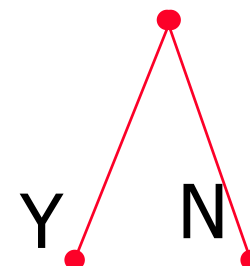
⌘ The method

☑ Divisive

Recursive, descendant

☑ Binary

Binary question

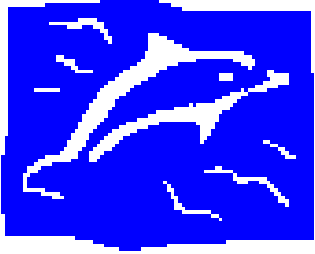


☑ Symbolic

Input: *symbolic data*

Output: *symbolic interpretation*

Clusters: *assertion object*



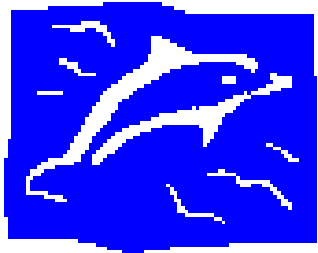
Method: DIV algorithm



⌘ The method

- ☒ A split of one cluster into two clusters
 - ☒ Binary question
 - ☒ Optimisation of evaluation criterion

→ Binary tree



Evaluation criterion (within-cluster inertia)



Additive criterion

$$P=(C_1, C_2, \dots, C_k)$$

$$W(P) = \sum_{C_k \in P} Q(C_k)$$

Q measures the quality of a cluster

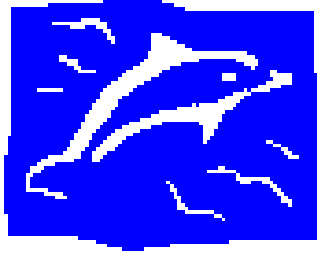
$$Q(C) = \frac{1}{2n_k} \sum_{\omega_i \in C_k} \sum_{\omega_j \in C_k} d^2(\omega_i, \omega_j)$$

n_k = number of individuals in C_k

d = distance or dissimilarity between symbolic objects (Hausdorff, KHI2)

Normalization:

- inverse of dispersion (symbolic variance)
- inverse of maximum deviation



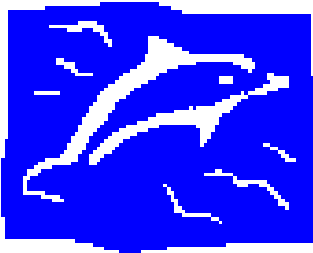
Evaluation criterion (within-cluster inertia)



$$\Delta(Q) = |Q(C) - Q(C_1) - Q(C_2)|$$

$\Delta(W)$: maximal

(C_1, C_2) : optimal subdivision of C



Choice of the cut value: s



⌘ Numerical or ordinal

- ☑ Order the value of the variables
- ☑ Choice s in the middle of 2 consecutive values

⌘ Interval

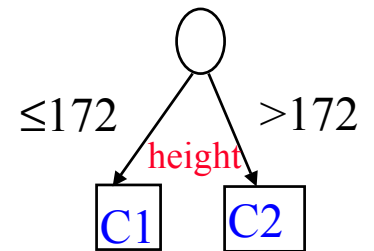
- ☑ Reduce the interval in a point: the centre
- ☑ Choice s : idem numerical method

⌘ Probabilistic

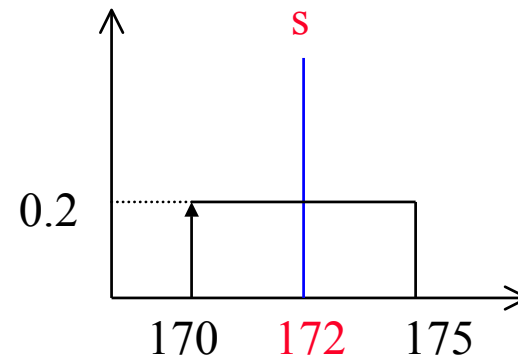
- ☑ On probabilistic distribution
- ☑ Choice s = mediane

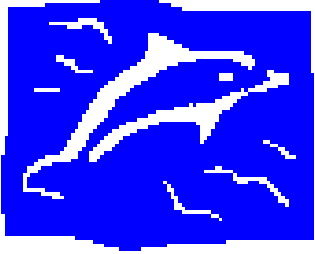
Individual ω de C :

height(ω) = $[170, 175]$



$\omega \in C2$

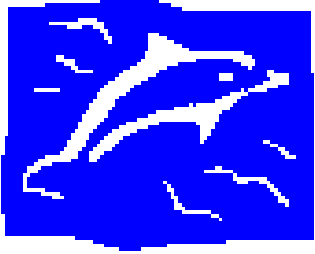




Method: DIV algorithm



- ⌘ **Step 1:** all objects in one cluster C
- ⌘ **Step 2:** divides successively each cluster C into smaller ones (C_1, C_2) according the **within-clusters inertia criterion**
 - ⊞ **Step 2.1:** for each variable y , find the cutting S which optimises $W(c) = q(c_1) + q(c_2)$
 - ⊞ **Step 2.2:** choose the variable y and the cutting S which optimises $W(C)$



Method: DIV algorithm



⌘ **Step 3:** split the cluster $C \longrightarrow (C1, C2)$
which maximises

$$\Delta(C) = |q(c) - q(c1) - q(c2)|$$

⌘ **End:** each division is carried out using a single variable and by separating objects possessing some specified values of this variable, from those lacking them



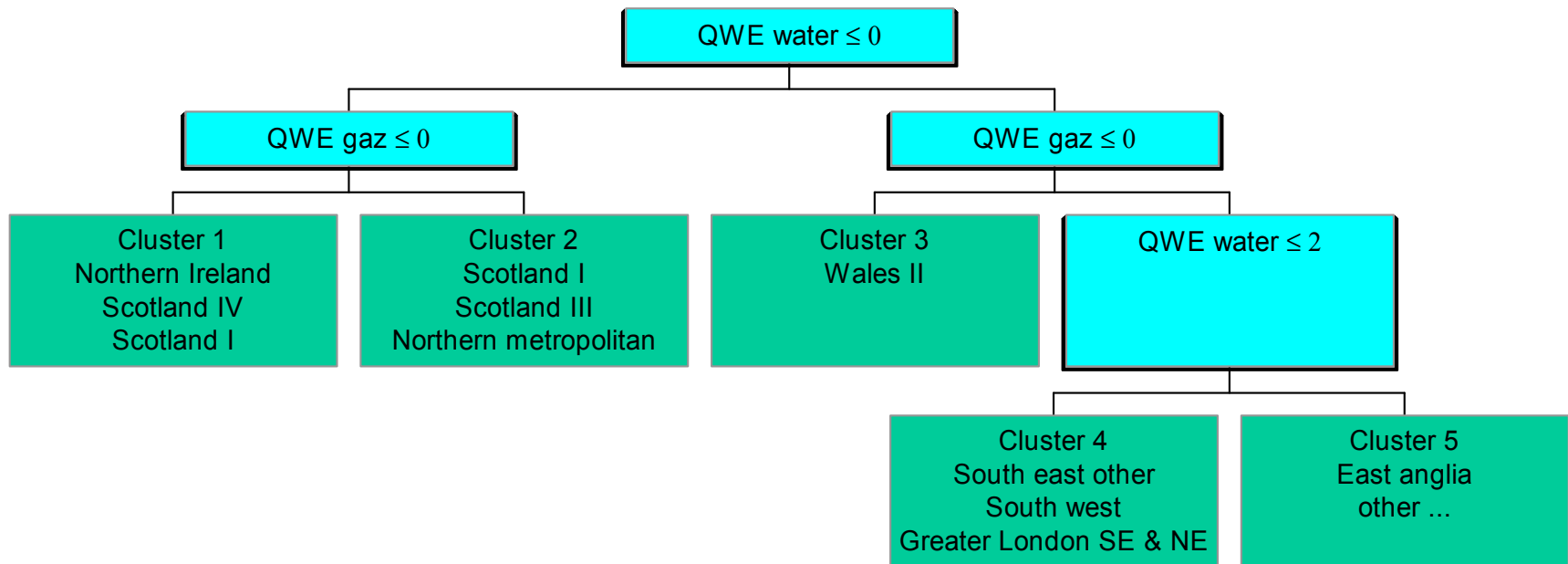
Output results

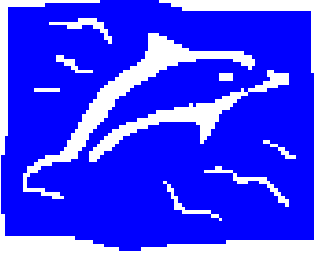
- ⌘ **Nested** partitions of the set of the individuals
- ⌘ Set of **assertion objects** describing the clusters of the concerned partition
- ⌘ Indexed hierarchy = **clustering** (binary, decision) **tree**
- ⌘ Each node represents a **class**
- ⌘ For each node or class, a **symbolic object** is associated



Output results

Clustering tree





Example: ONS census data file



⌘ 407 towns = 407 symbolic objects

⌘ 71 percentage interval variables

SO1 = location = Adur

And age0-4 = [5.80 : 6.00]

And white = [97.50 : 98.30]

And black = [0.19 : 0.21]

And Asian = [0.35 : 0.37]

And Nocar = [49.50 : 51.30]

And Primanu = [25.50 : 26.10]

And Wktravel = [91.50 : 92.20]

And class4-5 = [18.00 : 19.70]

Example: ONS SODAS chaining



SODAS file Chaining Options Window Help

Methods

Sodas procedures

(method name)

(method description)

SOE DIV STAT DKS DI

PCM FDA

LGDNouv.FIL

Chaining Model Method Window Help

LGD.SDS
d:\bases\011

BASE

Div
Divizive Classification

1 DIV

SOE
Symbolic Object Editor

2 SOE

Stat
Histogram, Elementary Statistics

3 STAT

END

Here, we need to use DB2SO to transform div output partition into SODAS symbolic object format

Example : the clustering tree on ONS census data



+---- Class 1 (Ng=97)

CLUSTER 1 : In these towns :

- minus 22% of inhabitants has no car
- minus 15% belongs to CLASS4-5

It seems to be a cluster of **privileged towns**

Towns of **NOCAR** percentage $\leq 22\%$



! ! !
! ! !
! ! !

!----3- [CLASS4_5 ≤ 15.045000]

! !

! +---- Class 4 (Nd=150)

! !

!----1- [NOCAR ≤ 22.235000]

! !

+---- Class 2 (Ng=69)



First cutting: variable **NOCAR**

! !----4- [PRIMMANU ≤ 22.630000]

! ! !

! ! +---- Class 5 (Nd=70)

! !

!----2- [WKTRAVEL ≤ 34.700001]

! !

+---- Class 3 (Nd=20)

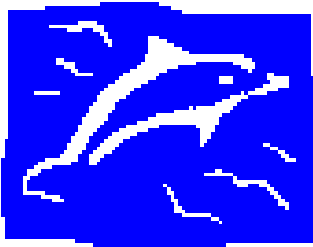
CLUSTER 3 : In these towns :

- more than **22%** of inhabitants has no car
- more than **35%** of inhabitants travel to work by public transport

It seems to be a cluster of **underprivileged towns**

Towns of **NOCAR** percentage $\geq 22\%$





Symbolic object description cluster 1/3



AGE04 = [4.55 : 8.01]
And AGE514 = [9.12 : 13.90]
" "

And WHITE = [73.41 : 99.57]
And BLACK = [0.08 : 4.24]
And ASIAN = [0.04 : 17.82]
" "

And ADT1_KID = [1.82 : 4.69]
And OWNEROCC = [68.75 : 91.34]
" "

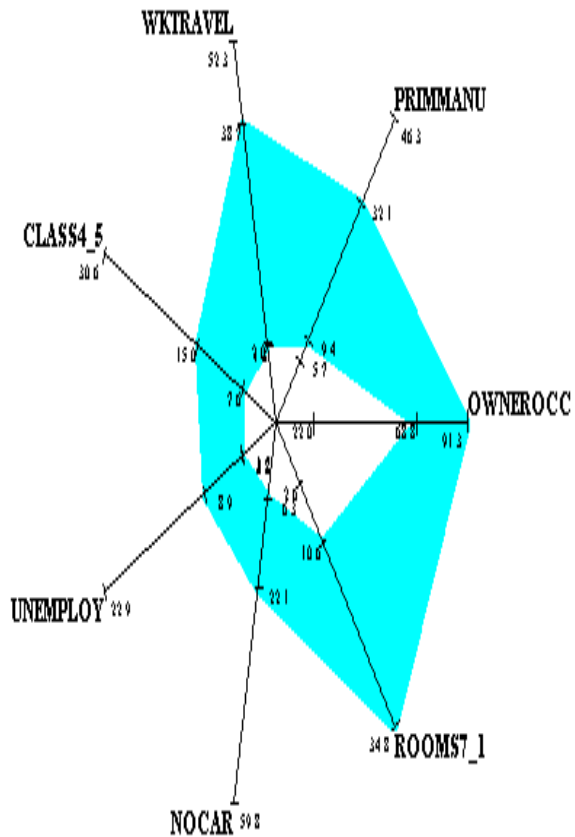
And ROOMS7_2 = [14.26 : 42.68]
And NOCAR = [6.25 : 22.15]
" "

And CLASS4_5 = [7.59 : 15.04]
And WKTRAVEL = [1.96 : 38.67]
And AGRWORK = [0.12 : 9.43]
And SERVWORK = [30.81 : 58.15]
And PRIMMANU = [9.42 : 32.11]

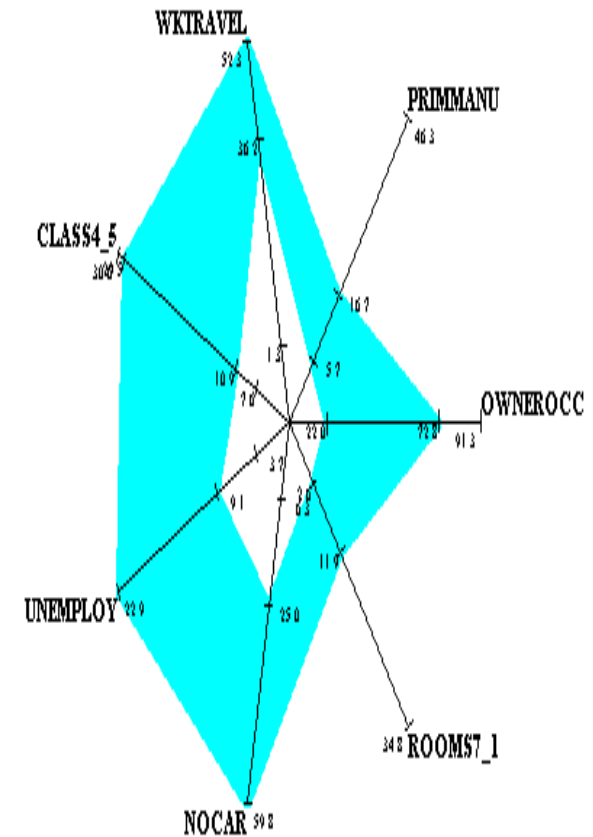
Visualisation of clusters (SOE) on clusters 1 and 3

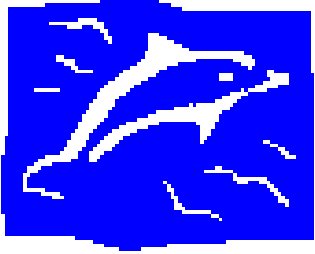


CLASS1/5



CLASS3/5

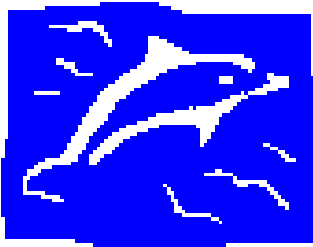




Conclusion on divisive clustering method



- ⌘ Symbolic data in input
- ⌘ Graphic visualisation of clusters
- ⌘ Symbolic and explicative representation of clusters



References



⌘ **M. Chavent**

« *A divisive and symbolic clustering method* » AMSDA'97

⌘ **M. Chavent, V. Stéphan**

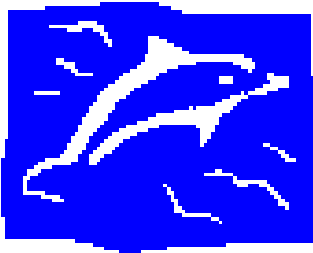
« *From generalization to clustering in the relational database context* »
KESDA'98

⌘ **M. Chavent**

« *Analyse des données symboliques, une méthode divisive de classification* » Thèse de l'Université Paris IX-Dauphine, 1997

⌘ **M. Chavent**

« *A monothetic clustering method* » pattern recognition letters 19, pp. 989-996,
1998



References

⌘ **M. Chavent.**

« *Criterion-Based Divisive Clustering for Symbolic Data* »,

Analysis of Symbolic Data, in Classification, Data Analysis and Knowledge organisation, éd. H.Bock, E. Diday, Springer-Verlag Edition, 1999.

⌘ **M. Touati, F. Goupil, E. Diday, J. Charlton.**

« *Le logiciel SODAS pour l'analyse de Données Symboliques : une application à des données du recensement du Royaume-Uni* »,

SFC99, 7èmes journées de la Société Francophone de Classification, Nancy, 15-17 Septembre 1999.

⌘ **M. Touati, F. Goupil, E. Diday, R. Moul.**

« *Processing CENSUS data from ONS* »

Analysis of Symbolic Data, in Classification, Data Analysis and Knowledge organisation, éd. H.Bock, E. Diday, Springer-Verlag Edition, 1999.