A divisive approach for clustering symbolic data

M. Chavent

Mathématiques Appliquées de Bordeaux (MAB),
Universités Bordeaux1 et 2

Plan :

1. Divisive clustering method

   - ✓ descendant hierarchical algorithm
   - ✓ classical or symbolic data

2. Application for clustering a set of categories

   - ✓ example of a set of species contaminated with mercury
   - ✓ comparison of numerical and symbolic approach for clustering the species

Introduction to divisive clustering

Hierarchical clustering

1. Starts with the whole dataset
2. At each step :
   - ✓Choice of the cluster to be split
   - ✓Split of this cluster
3. Stop :
   - ✓After K iterations or
   - ✓All clusters are singletons or have similar objects

Introduction to divisive clustering

How to split a cluster

- ✓ Complete enumeration (Edward and Cavalli-Sforza, 1965)
- ✓ Heuristics (Mac-Naughton-Smith, 1964, Chidananda Gowda et al. 1978)
- ✓ Criterion optimisation like split or diameter (Guénoche 1991, Wang et al. 1996)
- ✓ **Monothetic approach**
  - ✓Clustering : Williams & Lambert 1959
  - ✓Classification :Breiman et al. 1984 (CART) or Quinlan 1986 (ID3)

The proposed divisive clustering method

Takes classical or symbolic data as input

	height	weight
1	[174,176]	62
2	[160,163]	56
3	[183,185]	80

	height	weight
1	(small 0,3;large 0,7)	{very light,light}
2	(small 0,8;large 0,2)	{very light}
3	(small 0;large 1)	{heavy}

$X = \left(x_i^j \right)$ with $x_i^j =$

single value
interval of values
set of categories
probability distribution

- Definition of a dissimilarity measure
- Extension of the within-cluster inertia criterion

The proposed dissimilarity measure

For a numerical symbolic (interval) data matrix :

$$d^2(x_i, x_{i'}) = \sum_{j=1}^{p} d_H^2(x_i^j, x_{i'}^j)$$

- d_H is the **Hausdorff distance** between two intervals
- If all intervals are reduced to single points, it corresponds to the Euclidean distance

The normalized version of this distance :

$$d^2(x_i, x_{i'}) = \sum_{j=1}^{p} \frac{1}{v^j} d_H^2(x_i^j, x_{i'}^j) \qquad v^j = \frac{1}{n^2} \sum_{i=1}^{n-1} \sum_{i'>i}^{n} d_H^2(x_i^j, x_{i'}^j)$$

The proposed dissimilarity measure

From a symbolic categorical data matrix

	height		weight		
	small	large	very light	light	heavy
1	0,3	0,7	0,5	0,5	0
2	0,8	0,2	1	0	0
3	0	1	0	0	1

→

	1	…	j	…	m
1					
…					
i			xij		
…					
n					

np

$$d^2(x_i, x_{i'}) = \sum_{j=1}^{p} d^2_{\phi 2}(x_i^j, x_{i'}^j)$$

- $d_{\Phi 2}$ is the $\Phi 2$ **distance** between two frequency vectors

The extension of the within-cluster inertia

Let $D=\left(d_{ii'}\right)$ the distance matrix

- the extension of the inertia of a cluster is :

$$I(C_k)=\sum_{i\in C_k}\sum_{i'\in C_k}\frac{1}{2nn_k}d_{ii'}^2$$

- the extension of the within-clusters inertia of a partition in K clusters is :

$$W(P)=\sum_{k=1}^{K}I(C_k)$$

The proposed divisive clustering method

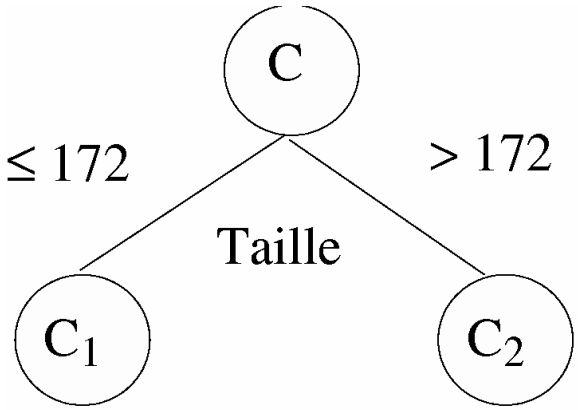
Uses a variable and a binary transformation of this variable to split a cluster

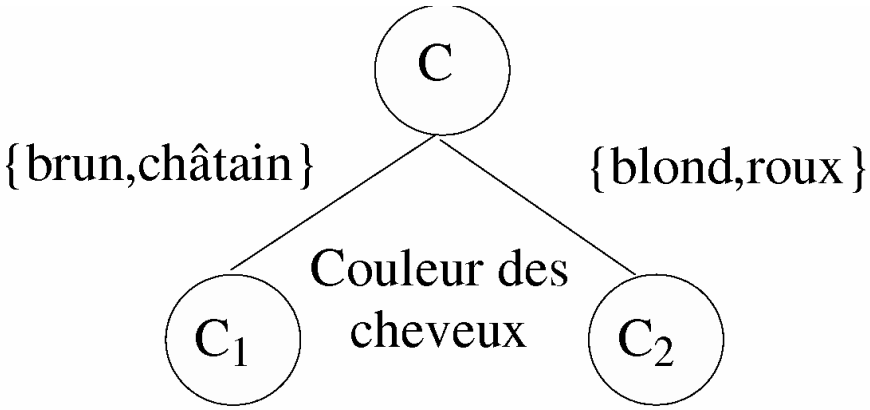
- extension of the definition of binary question to the case of symbolic data

Choose the best biparition among all possible bipartition in 2 clusters

- how many possible bipartition to evaluate ?

Binary question and symbolic data





How many possible bipartitions

If the variable Y_j is quantitative (classical or interval) there is at most $z_j=n_k-1$ bipartitions induced by this variable

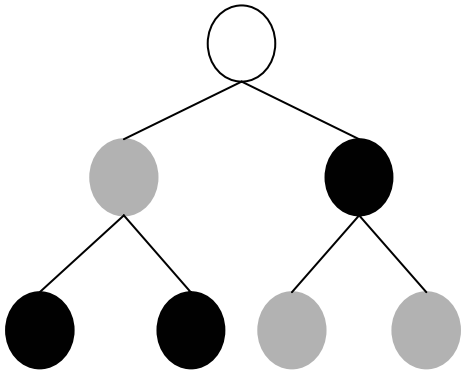
If the variable Y_j is qualitative there is at most :

- $z_j=m_j-1$ bipartitions induced for ordered categories
- $z_j=2^{mj-1}-1$ bipartitions induced for non ordered categories

⇨ $z_1+\ldots+z_p$ biparitions to evaluate

The proposed divisive clustering method

Choose at each step the cluster to be split such that the new partition has minimum within inertia ?



$$\Delta(C_k) = Q(C_k) - Q(C_k^1) - Q(C_k^2)$$

Stops after L iterations given as input by the user and the output is a hierarchy indexed by Δ, which is isotonic

Application for clustering a set of categories

FISH	SPECIES	Ln(liver/Muscle)	…	Ln(Stomach/Muscle)
1	Ageneiosus brevifili	-0.12	…	-1.77
	Ageneiosus brevifili	-0.30	…	-1.22
…	…	…	…	…
	Doras micropoeus	0.8	…	-0.89
67	Doras micropoeus	1.34	…	-1.45

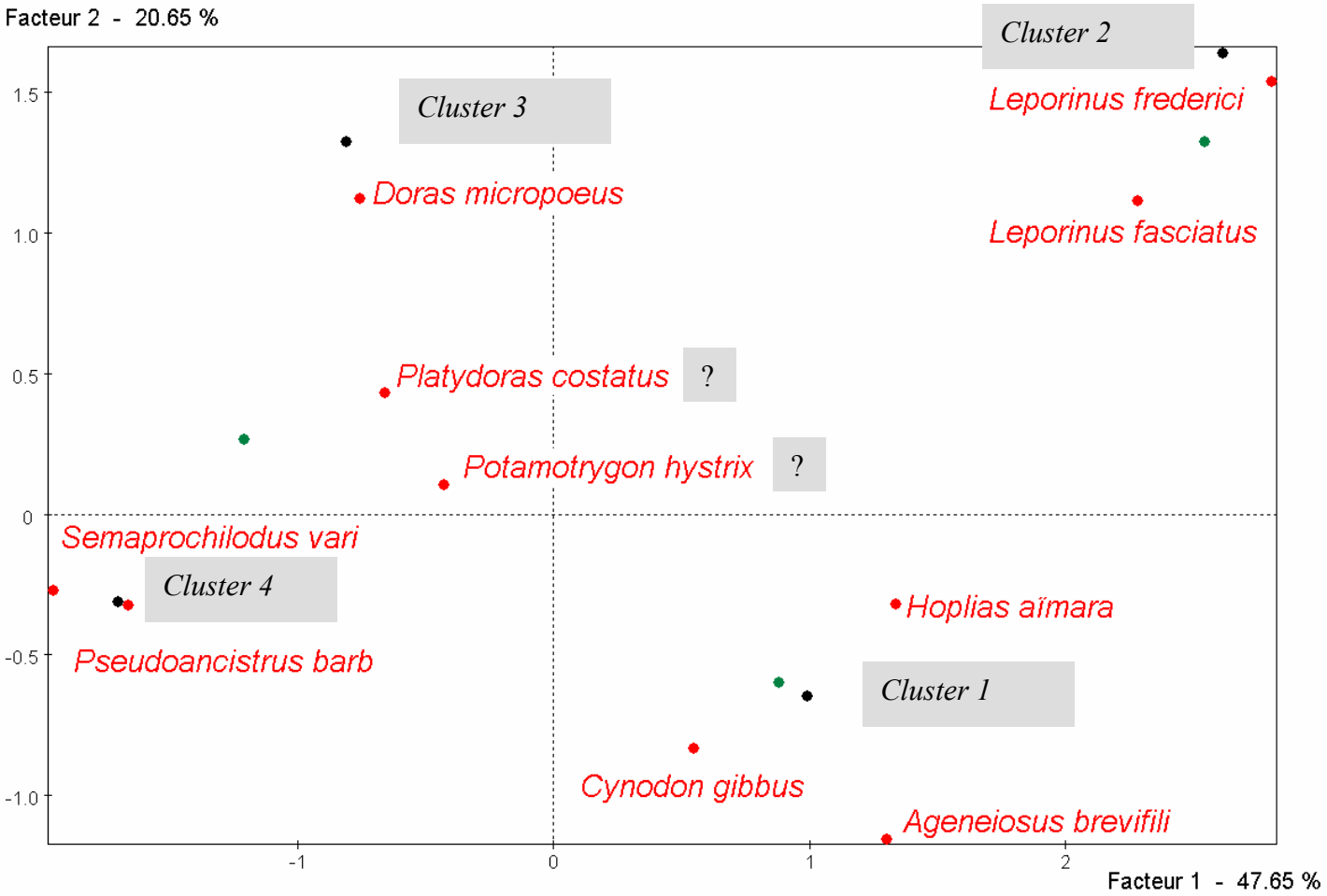
⇨ Clustering the 10 categories of SPECIES

First numerical treatment

Clustering of the 67 fish :Ward hierarchical clustering on the 67 fish $\times$ 5 concentration variables

	Cluster1	Cluster2	Cluster3	Cluster 4
Ageneiosus brevifili	100	0	0	0
Cynodon gibbus	100	0	0	0
Hoplias aimara	100	0	0	0
Doras micropoeus	0	0	100	0
Leporinus fasciatus	33.33	66.67	0	0
Leporinus frederici	0	100	0	0
Pseudoancistrus barbatus	14.29	0	0	85.71
Semaprochilodus varii	0	0	0	100
Platydoras costatus	20	0	40	40
Potamotrygon hystrix	50	0	25	25

First numerical treatment



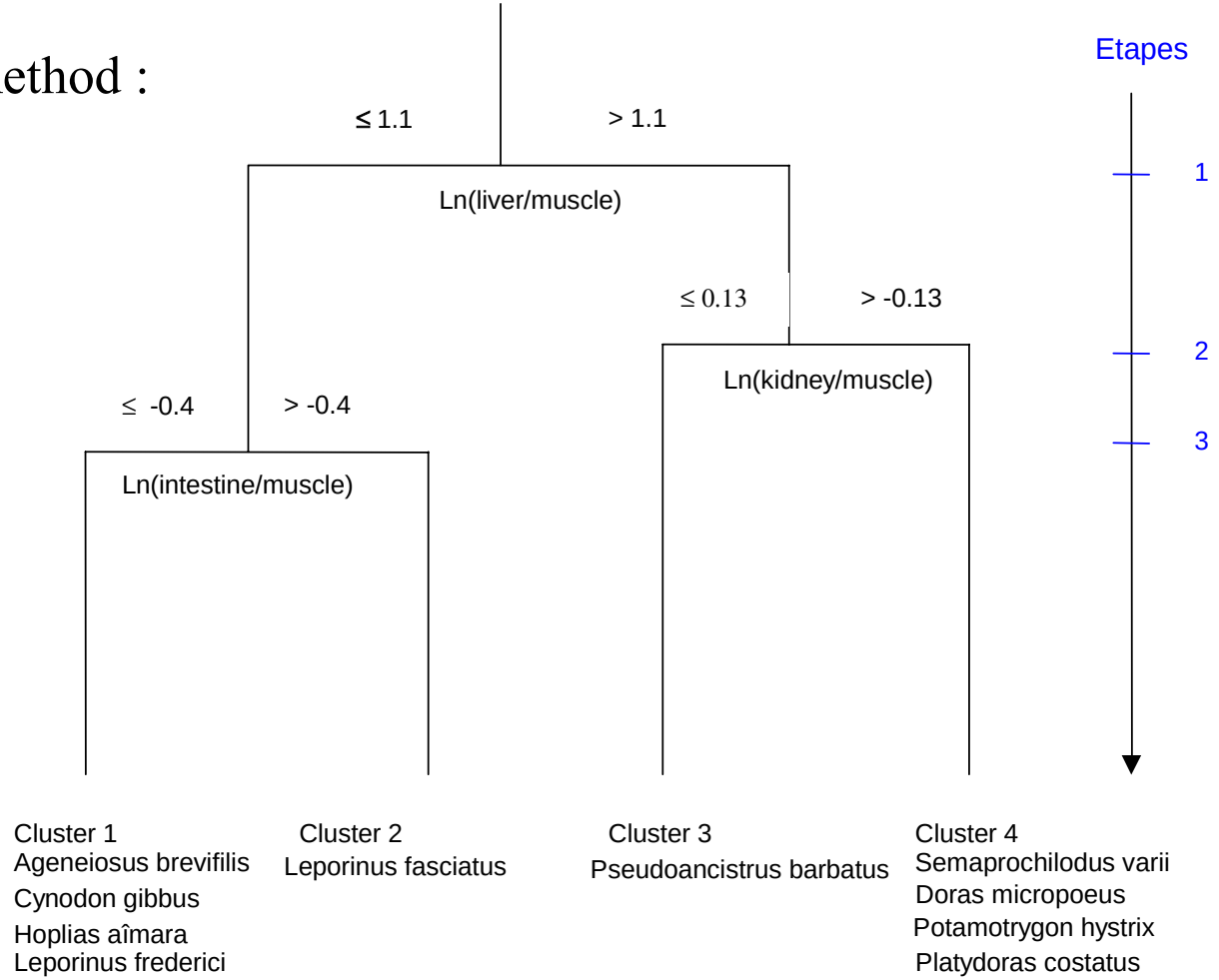
Second numerical treatment

Clustering of the 10 species on a classical numerical data set…

NUM	SPECIES	Ln(liver/Muscle)	…	Ln(Stomach/Muscle)
1	Ageneiosus brevifili	-0.38	…	-1.17
…	…	…	…	…
10	Doras micropoeus	1.72	…	-0.84

Second numerical treatment

… with the DIV method :



Symbolic treatment

Clustering of the 10 species on a symbolic data set…

NUM	SPECIES	Ln(liver/Muscle)	…	Ln(Stomach/Muscle)
1	Ageneiosus brevifili	[-0.8;-0.08]	…	[-1.48;-0.59]
…	…	…	…	…
10	Doras micropoeus	[1.34;2.12]	…	[-1.45;-0.24]

Symbolic treatment

… with the DIV method :