Hierarchical and Pyramidal Clustering for Symbolic Data

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

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

U.PORTO
FEP FACULDADE DE ECONOMIA UNIVERSIDADE DO PORTO
LIAAD INESCPORTO LA

ECI 2015
20 AL 25 DE JULIO

Outline

- Clustering structures
  - From the hierarchical to the pyramidal model
- Symbolic Clustering
  - The generalization procedure
  - The generality degree
  - The clustering algorithm
  - The *HIPYR* Module of *SODAS*

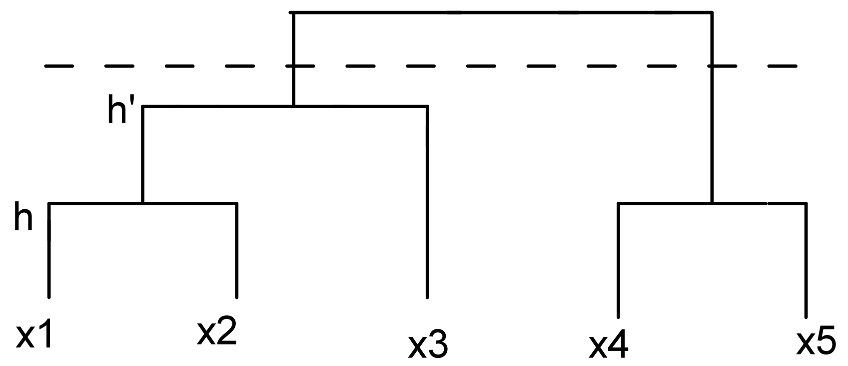
Clustering structures

Hierarchical Model: set of nested partitions

Let S be the observations set (the set being clustered)

Hierarchy on S:
Family H on non-empty subsets of S such that

- $S \in H$
- $\forall\, s \in S\,,\, \{ s \} \in H$
- $\forall\, h, h' \in H, \quad h \cap h' = \emptyset$ or
  $h \subseteq h'$ or $h' \subseteq h$



Clustering structures

Pyramidal model:

Compatibility between a dissimilarity and an order

S - the observations set (the set being clustered)

d - dissimilarity index on S

θ - linear order on S

d and θ are COMPATIBLE iff, for any ordered triplet,

$$s_i \ \theta \ s_j \ \theta \ s_k$$

$$d(s_i , s_k) \geq \text{Max} \{ d(s_i , s_j) , d(s_j , s_k) \}$$

Clustering structures

Pyramid P on S

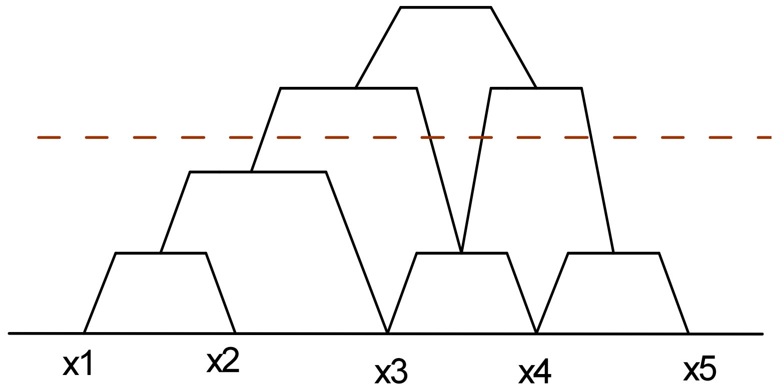
Family P on non-empty subsets of S such that :

- $S \in P$
- $\forall\ s \in S\ ,\ \{\ s\ \} \in P$
- $\forall\ p, p' \in P,\quad p \cap p' = \emptyset\quad$ or $p \cap p' \in P$
- There exists a linear order θ : every element of P is an interval of θ

Pyramidal model :

→ Clustering

→ Seriation



Hierarchy : nested partitions

Pyramid : nested overlappings

Clustering structures

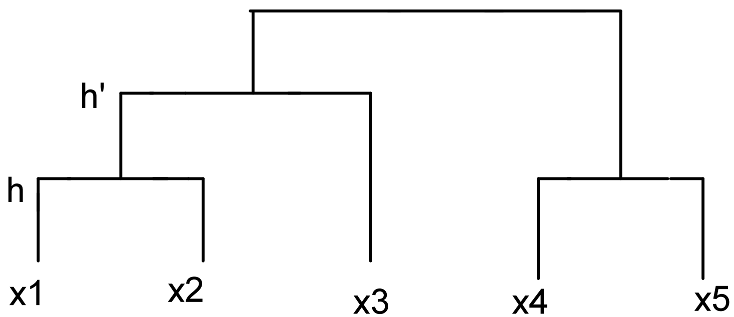
Successor and Predecessor

C – Hierarchy or Pyramid

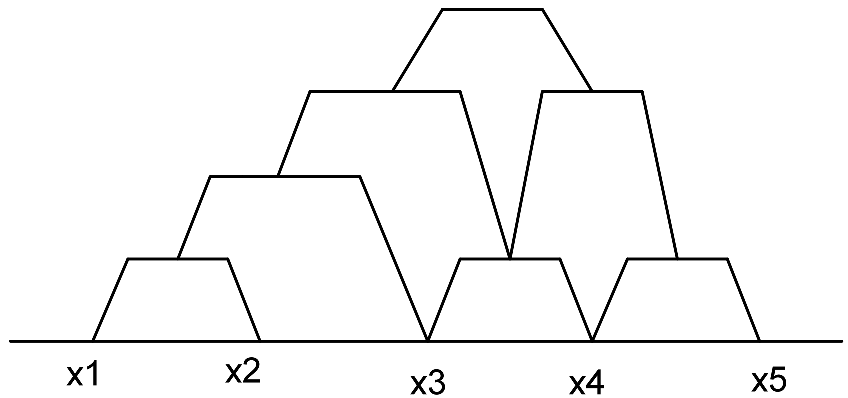
$p \in P$ SUCCESSOR of $p' \in C$ if

1) $p \subseteq p'$ 2) $\neg\exists\ p'' \in C : p \subseteq p'' \subseteq p'$

p' is a PREDECESSOR of p



<u>Hierarchy</u> : Each cluster has at most ONE predecessor



<u>Pyramid</u> : Each cluster has at most TWO predecessors

Clustering structures

Indexed Hierarchy and Indexed Pyramid

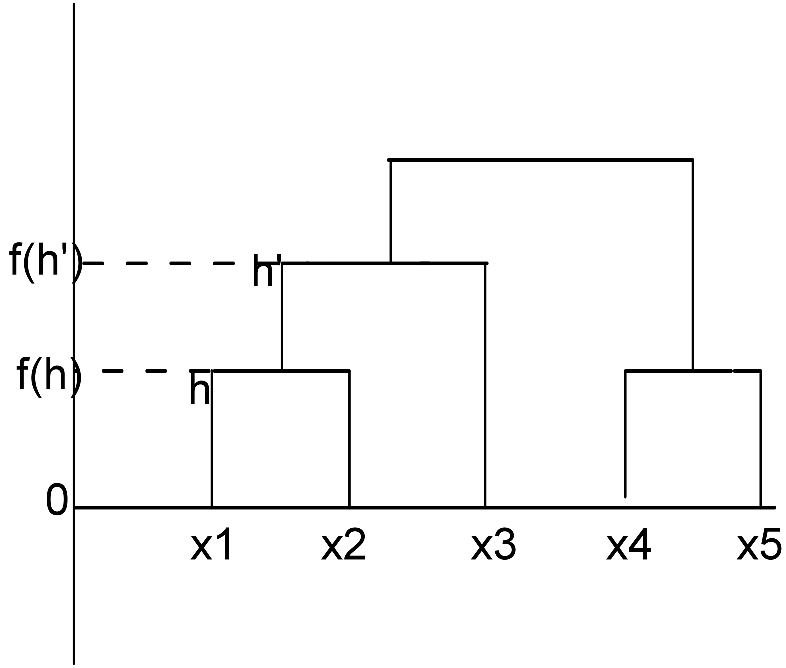
(C, f) with

C – Hierarchy or Pyramid

$f : C \rightarrow \mathbb{R}^+$

a) $f(h) = 0 \Leftrightarrow \#h = 1$

b) $h \subseteq h' \Rightarrow f(h) \leq f(h')$



<u>Pyramid Indexed in the Broad Sense</u> :

$f(p) = f(p')$ with $p \subset p'$ and $p \neq p' \Rightarrow$

$\exists\ p_1 \neq p,\ p_2 \neq p$ such that $p = p_1 \cap p_2$

Pyramidal (Robinsonian) index

Dissimilarity index d such that :

a) $d(x , y) = 0 \Rightarrow x = y$

b) there exists an order θ on S such that

$$\forall\, s_i , s_j , s_k \in S,\quad s_i \,\theta\, s_j \,\theta\, s_k \Rightarrow d(s_i, s_k) \geq \max \{d(s_i, s_j) , d(s_j, s_k) \}$$

Pyramidal (Robinsonian) index

$d(s_i, s_j)$ = height of the smallest cluster containing s_i and s_j

Johnson-Benzécri Theorem :

Bijection between indexed hierarchies and ultrametric dissimilarities

Hierarchy : d is an ultrametric dissimilarity

Theorem :

Bijection between pyramids indexed in the broad sense and pyramidal (robinsonian) indices

Pyramid : d is a pyramidal index

The matrix of d ordered according to θ is Robinson

Clustering structures

Ascending clustering algorithm

Starting with the one element clusters,
merge at each step the MERGEABLE clusters
for which the dissimilarity (aggregation index) is MINIMUM

Mergeable clusters :

- if the structure is a hierarchy :
- none of them has been aggregated before ;

- if the structure is a pyramid :
- none of them has been aggregated twice, and
- there is a total order θ on S such that the new and all previously formed clusters are intervals of θ.

Clustering structures

Aggregation Indices :

- Complete Linkage (Maximum Dissimilarity)
- Single Linkage (Minimum Dissimilarity)
- Mean Linkage (Average Dissimilarity)
- Diameter
- Ward (Inertia Increase)

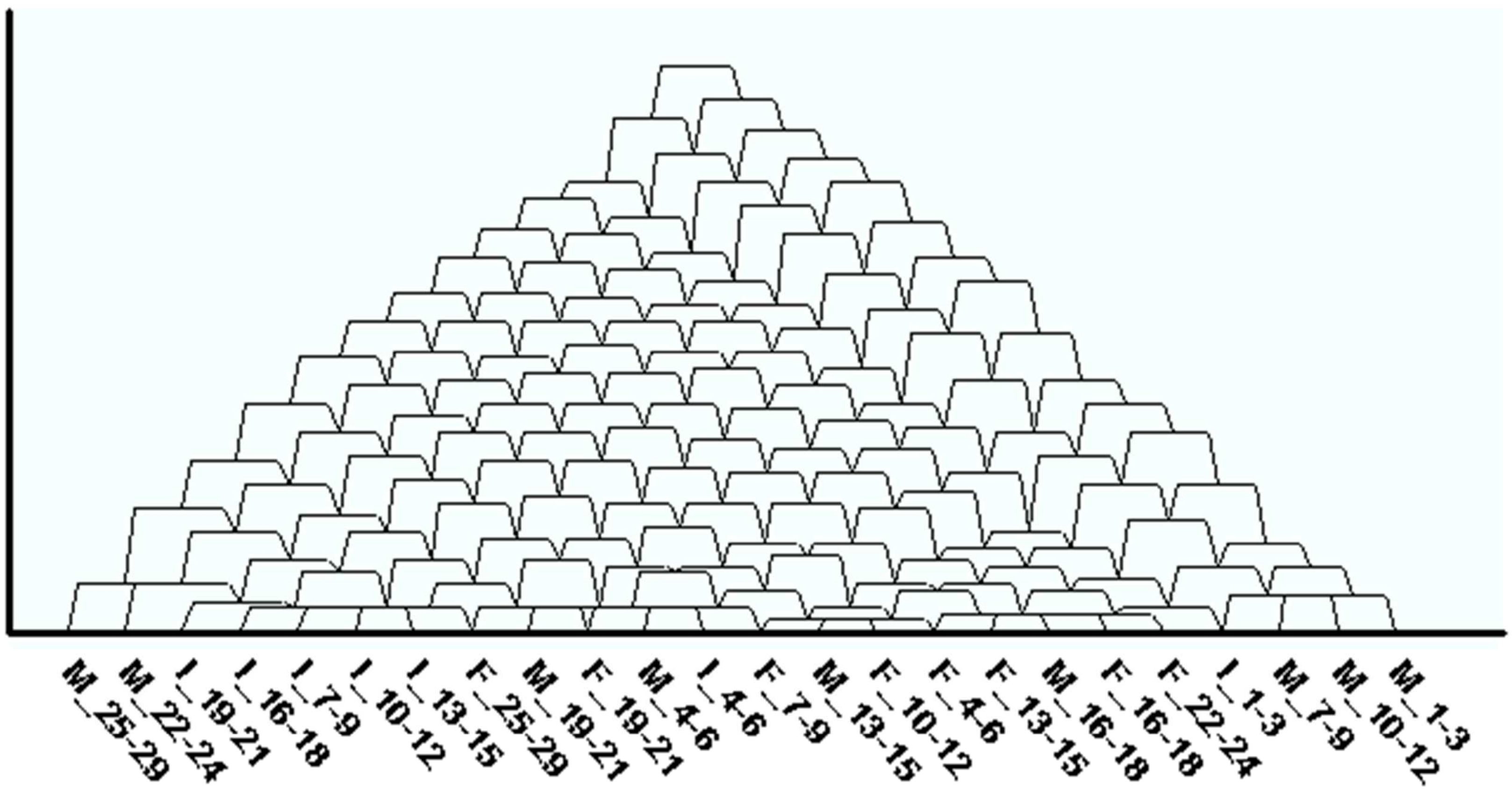
 ...

→ Lance & Williams recursive formula; generalized to pyramids

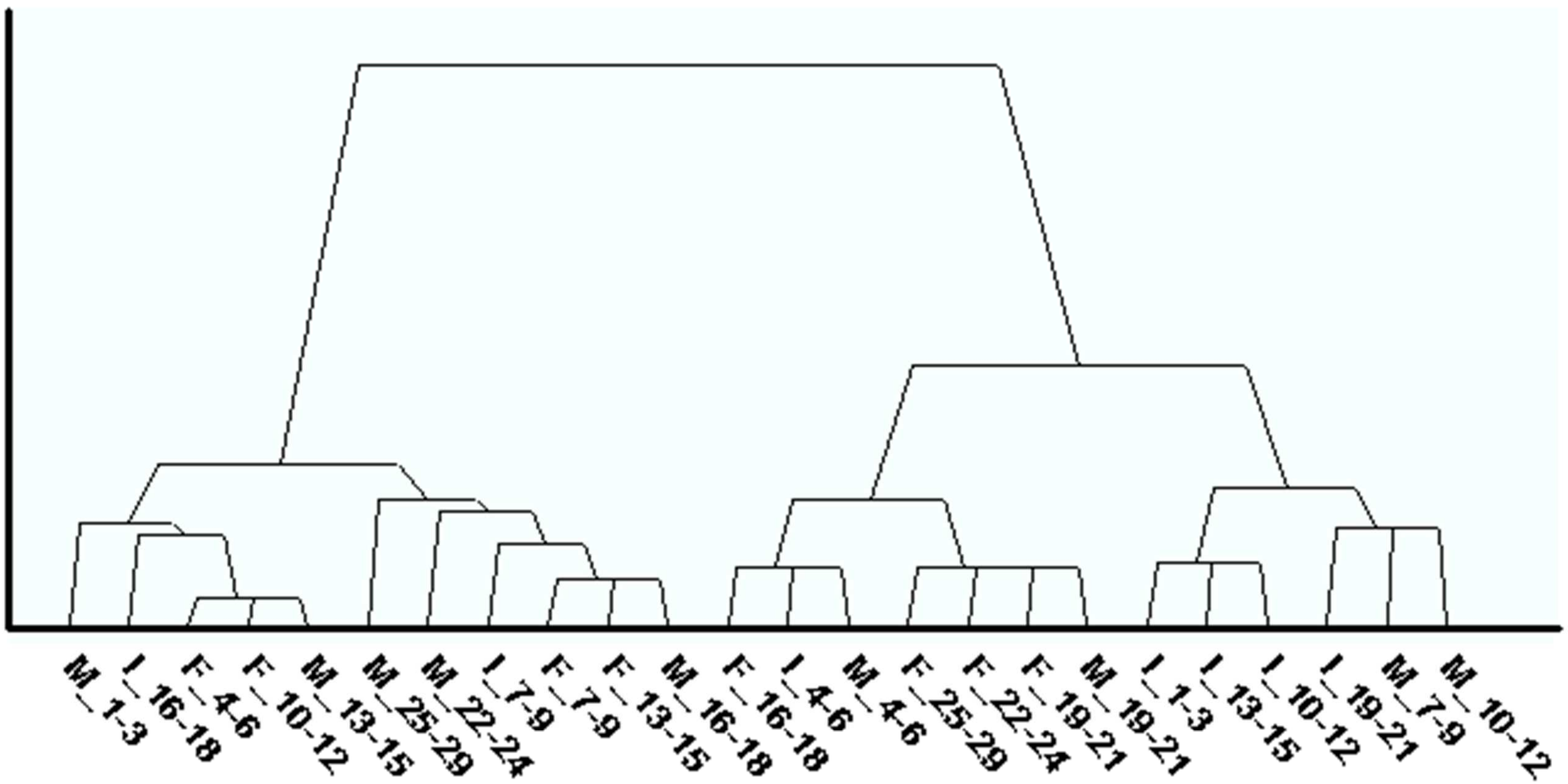
Abalone data

	LENGTH	DIAMETER	HEIGHT	WHOLE_WEIGHT	SHUCKED_WEIGHT	VISCERA_WEIGHT	SHELL_WEIGHT
F_4-6	[0.28 : 0.66]	[0.19 : 0.47]	[0.07 : 0.18]	[0.08 : 1.37]	[0.03 : 0.64]	[0.02 : 0.29]	[0.03 : 0.34]
F_7-9	[0.31 : 0.75]	[0.22 : 0.58]	[0.01 : 1.13]	[0.15 : 2.25]	[0.06 : 1.16]	[0.03 : 0.45]	[0.05 : 0.56]
F_10-12	[0.34 : 0.78]	[0.26 : 0.63]	[0.06 : 0.23]	[0.20 : 2.66]	[0.07 : 1.49]	[0.04 : 0.53]	[0.07 : 0.73]
F_13-15	[0.39 : 0.81]	[0.30 : 0.65]	[0.10 : 0.25]	[0.26 : 2.51]	[0.11 : 1.23]	[0.05 : 0.52]	[0.09 : 0.80]
F_16-18	[0.40 : 0.75]	[0.31 : 0.60]	[0.10 : 0.24]	[0.35 : 2.20]	[0.12 : 0.84]	[0.09 : 0.48]	[0.12 : 1.00]
F_22-24	[0.45 : 0.80]	[0.38 : 0.63]	[0.14 : 0.22]	[0.64 : 2.53]	[0.16 : 0.93]	[0.11 : 0.59]	[0.24 : 0.71]
F_19-21	[0.49 : 0.73]	[0.37 : 0.58]	[0.13 : 0.21]	[0.68 : 2.12]	[0.17 : 0.81]	[0.13 : 0.45]	[0.20 : 0.85]
F_25-29	[0.55 : 0.70]	[0.47 : 0.58]	[0.18 : 0.22]	[1.21 : 1.81]	[0.32 : 0.71]	[0.20 : 0.32]	[0.47 : 0.52]
I_1-3	[0.08 : 0.24]	[0.05 : 0.17]	[0.01 : 0.06]	[0.00 : 0.07]	[0.00 : 0.03]	[0.00 : 0.01]	[0.00 : 0.02]
I_4-6	[0.13 : 0.58]	[0.09 : 0.45]	[0.00 : 0.15]	[0.01 : 0.89]	[0.00 : 0.50]	[0.00 : 0.19]	[0.00 : 0.35]
I_7-9	[0.26 : 0.67]	[0.19 : 0.50]	[0.00 : 0.19]	[0.08 : 1.30]	[0.03 : 0.60]	[0.01 : 0.32]	[0.03 : 0.39]
I_13-15	[0.32 : 0.66]	[0.25 : 0.52]	[0.08 : 0.19]	[0.16 : 1.69]	[0.06 : 0.71]	[0.03 : 0.40]	[0.05 : 0.42]
I_10-12	[0.34 : 0.73]	[0.26 : 0.55]	[0.09 : 0.22]	[0.17 : 2.05]	[0.07 : 0.77]	[0.02 : 0.44]	[0.06 : 0.65]
I_16-18	[0.44 : 0.65]	[0.33 : 0.52]	[0.13 : 0.20]	[0.44 : 1.63]	[0.16 : 0.63]	[0.07 : 0.34]	[0.13 : 0.53]
I_19-21	[0.45 : 0.58]	[0.35 : 0.44]	[0.12 : 0.19]	[0.41 : 1.18]	[0.11 : 0.39]	[0.07 : 0.22]	[0.16 : 0.31]
M_1-3	[0.16 : 0.21]	[0.11 : 0.15]	[0.04 : 0.05]	[0.02 : 0.04]	[0.01 : 0.02]	[0.00 : 0.01]	[0.00 : 0.01]
M_4-6	[0.16 : 0.53]	[0.12 : 0.41]	[0.03 : 0.16]	[0.02 : 0.81]	[0.01 : 0.32]	[0.00 : 0.15]	[0.00 : 0.35]
M_7-9	[0.20 : 0.73]	[0.16 : 0.57]	[0.05 : 0.20]	[0.04 : 2.33]	[0.02 : 1.25]	[0.01 : 0.54]	[0.02 : 0.52]
M_10-12	[0.29 : 0.78]	[0.22 : 0.63]	[0.06 : 0.51]	[0.12 : 2.78]	[0.04 : 1.35]	[0.03 : 0.76]	[0.04 : 0.68]
M_13-15	[0.35 : 0.76]	[0.25 : 0.61]	[0.09 : 0.24]	[0.21 : 2.55]	[0.10 : 1.35]	[0.05 : 0.57]	[0.06 : 0.76]
M_16-18	[0.43 : 0.77]	[0.31 : 0.60]	[0.12 : 0.24]	[0.35 : 2.83]	[0.11 : 1.15]	[0.06 : 0.48]	[0.13 : 0.90]
M_19-21	[0.49 : 0.74]	[0.38 : 0.59]	[0.13 : 0.23]	[0.57 : 2.13]	[0.22 : 0.87]	[0.12 : 0.49]	[0.17 : 0.58]
M_22-24	[0.51 : 0.69]	[0.40 : 0.54]	[0.14 : 0.22]	[0.75 : 1.84]	[0.25 : 0.74]	[0.13 : 0.35]	[0.25 : 0.58]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

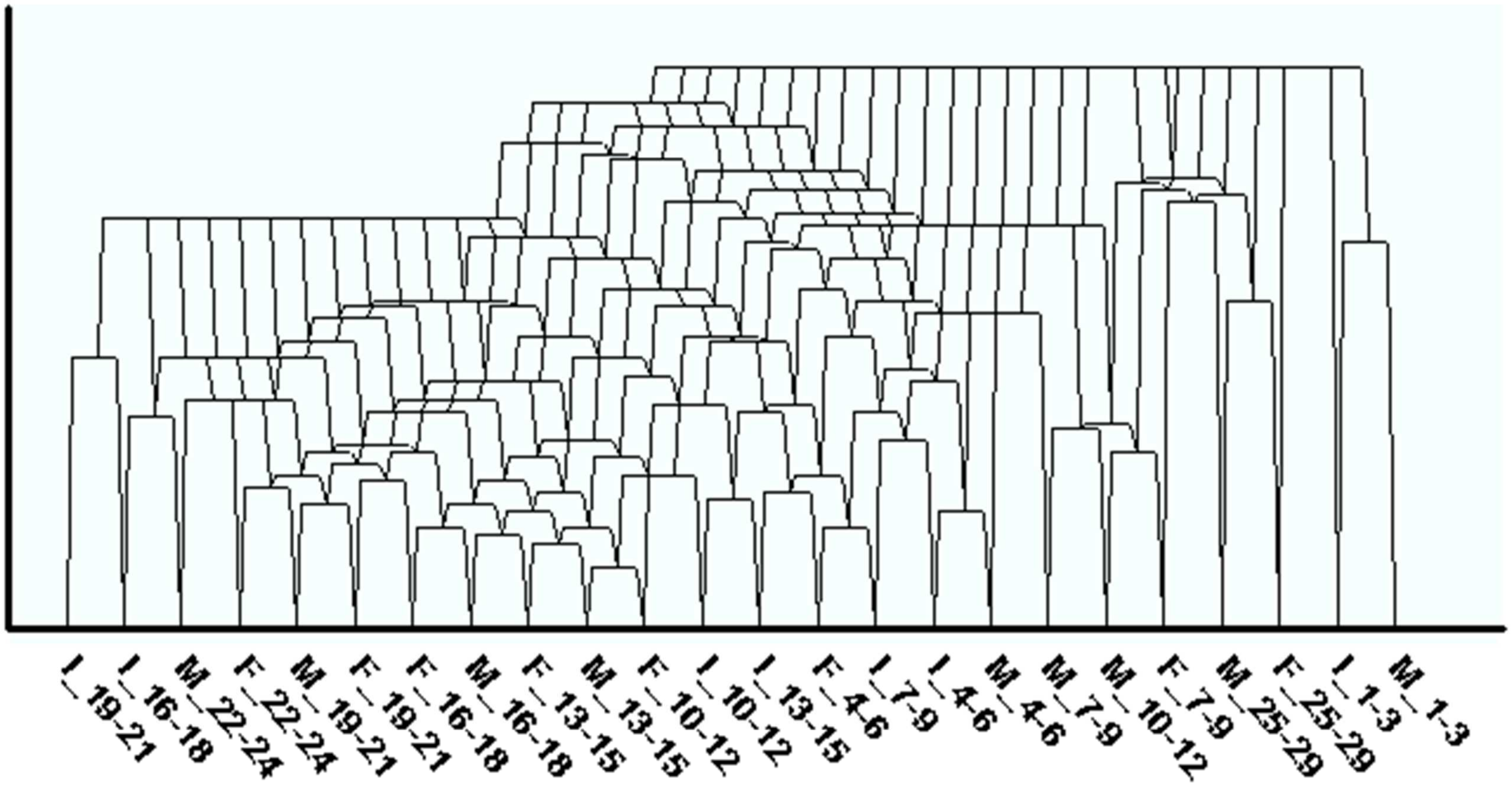
Abalone data: Mean linkage pyramid



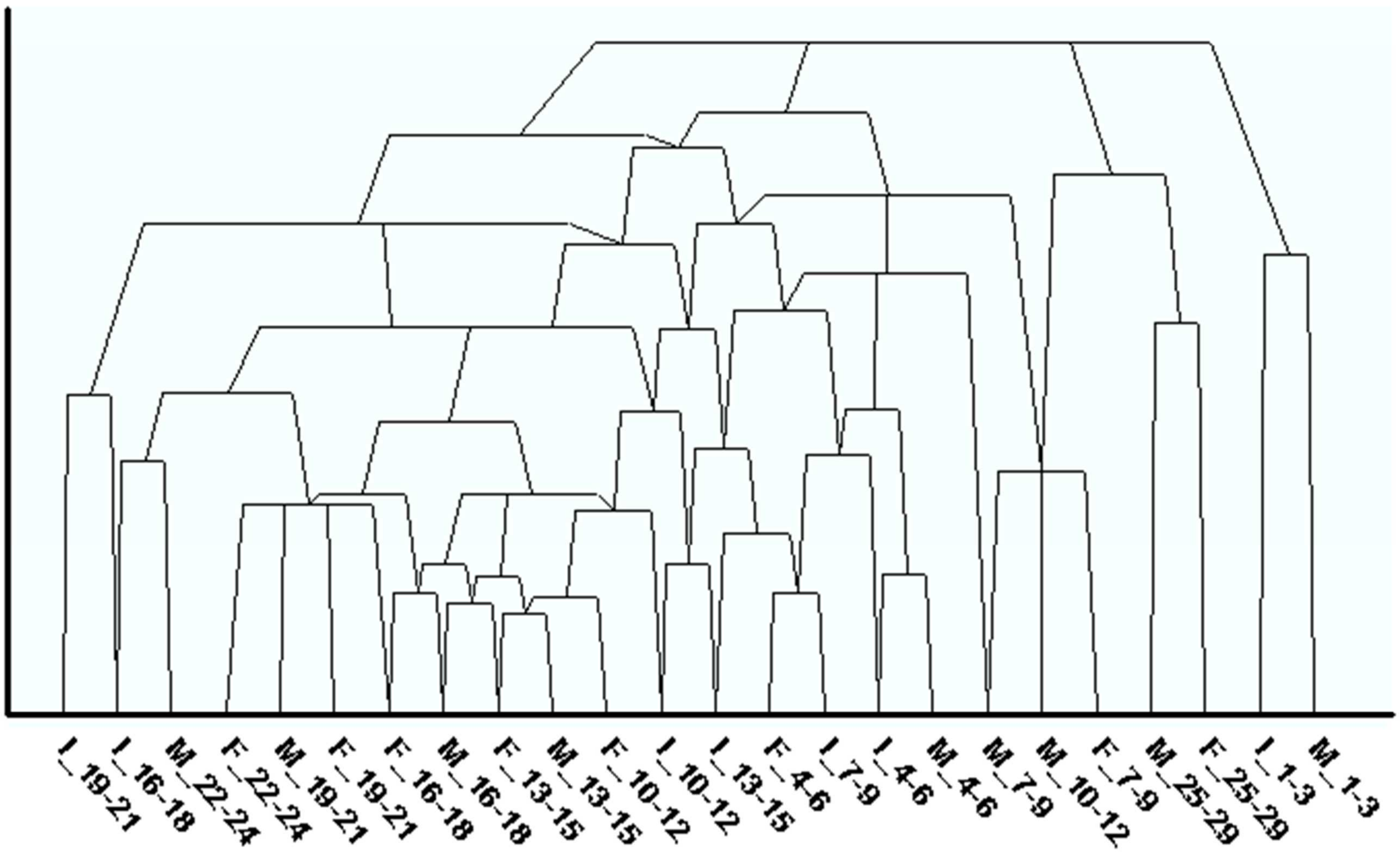
Abalone data:Mean linkage hierarchy



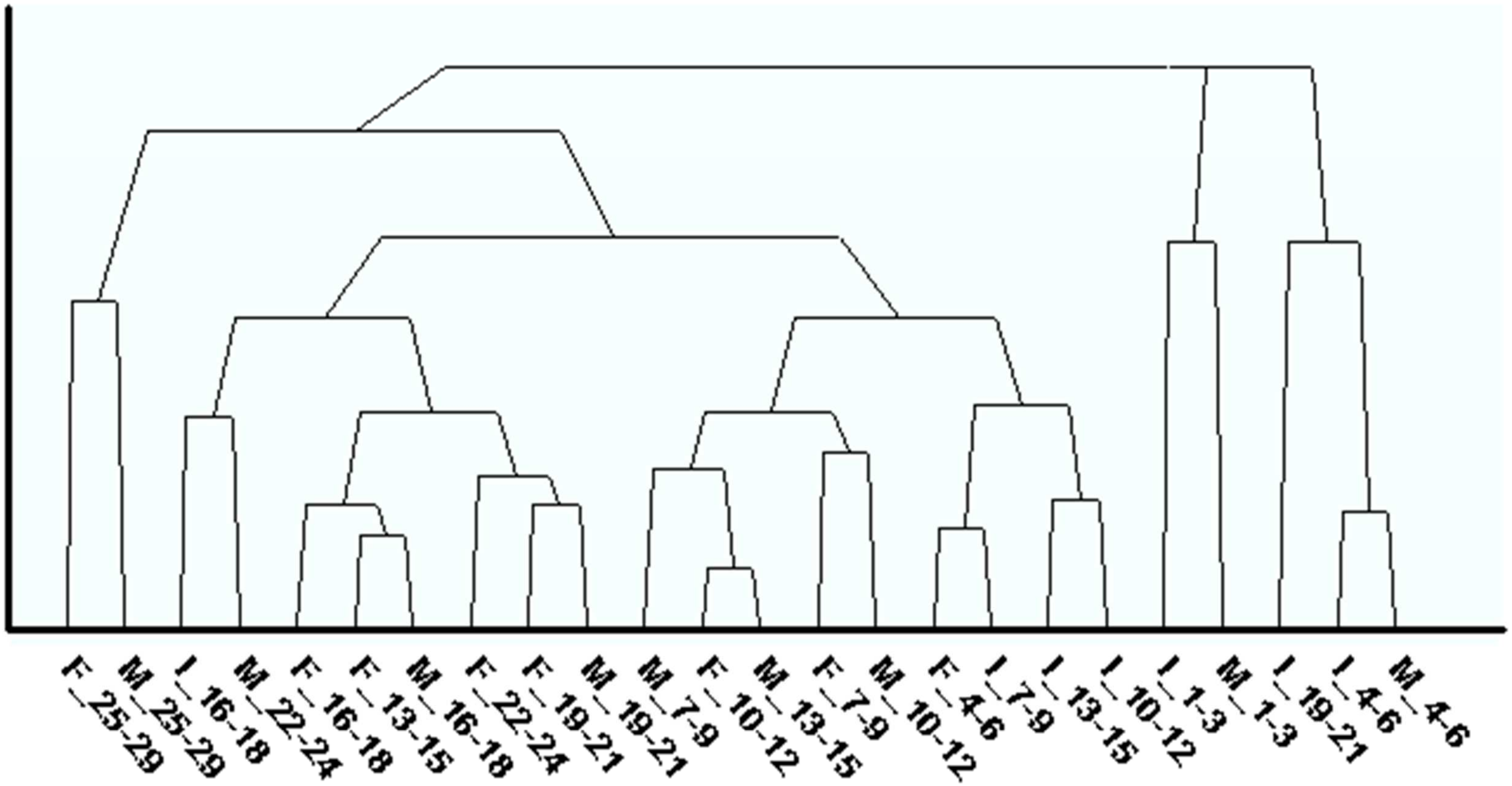
Abalone data: Complete linkage pyramid



Abalone data:
Complete linkage pyramid 10% pruned



Abalone data: Complete linkage hierarchy



From classical to symbolic data

Description: p-tuple $(d_1, \ldots, d_p)$, $d_j \in B_j$

Description space : $B = B_1 \times \ldots \times B_p$

Example:

([1000 ,15000] , {drinks (1/4), food (1/2), clothing (1/4)} ,

{Electron, Visa, Mastercard})

Let $S = \{s_1, \ldots, s_n\}$ the observed set

Then : $Y_j(s_i) \in B_j$ $j=1,\ldots, p$, $i=1,\ldots, n$

The data array consists on n descriptions, one for each $s_i \in S$:

$$(Y_1(s_i), \ldots, Y_p(s_i)) \quad , \quad i=1,\ldots, n$$

Extent and Intent

Extent of a description $d = (d_1, \ldots, d_p) \in B$,

Ext (d) : the set of elements $s \in S$ for which

Y_j (s) verifies d_j , $j=1,\ldots, p$

Intent of a subset $C \subseteq S$, Int(C) :

the description $d = (d_1, \ldots, d_p) \in B$

such that d_j is the minimal element in B_j $(j=1,\ldots, p)$

fulfilling the condition Y_j (s) verifies d_j $\forall s \in C$

Example :

	age	salary
s_1	[20 , 45]	[1000 , 3000]
s_2	[35 , 40]	[1200 , 3500]
s_3	[25 , 45]	[2000 , 4000]
s_4	[30 , 50]	[2000 , 3200]

d = ([20 , 45]] , [1000 , 4000])

$\text{Ext}(d) = \{ s \in S : \text{age}(s) \subseteq [20, 45]] \wedge \text{salary}(s) \subseteq [1000, 4000] \}$

$\text{Ext}(d) = \{ s_1, s_2, s_3 \}$

Concept

A concept is a pair (C, d) such that

- C is a subset of S
- d is a description, $d \in B$
- d is the intent of C : $Int(C) = d$
- C is the extent of d in E: $Ext_S(d) = C$

Example :

	age	salary
s_1	[20 , 45]	[1000 , 3000]
s_2	[35 , 40]	[1200 , 3500]
s_3	[25 , 45]	[2000 , 4000]
s_4	[30 , 50]	[2000 , 3200]

Int ({ s_1 , s_2 , s_3 }) = d = ([20 , 45] , [1000 , 4000])

Ext (d) = { s_1 , s_2, , s_3 }

Int (Ext (d)) = d

({ s_1 , s_2 , s_3 } , d) is a concept

Symbolic clustering

Objective :

Given a symbolic data array

build an hierarchical / pyramidal clustering

such that each cluster is a concept, i.e., a pair

$$\begin{cases} \text{EXTENSION - its members} \\ \text{INTENSION - its descriptions} \end{cases}$$

→ Each cluster has an automatic representation in terms of the descriptive variables

Symbolic clustering

Conceptual clustering methods require:

- Generalization Operator

  $C \subseteq C'$

  d' (representing C') is more general than

  d (representing C)

- Generality degree measure

Symbolic clustering: Generalisation

→ For a given Extent operator :

d is more general than d' if

the extent of d contains the extent of d'

d' is more specific than d

Generalisation of two descriptions d and d' :

determining d'' : d'' is more general than both d and d',

$Ext(d'') \supseteq Ext(d)$ and $Ext(d'') \supseteq Ext(d')$

This procedure differs according to the variable type

Generalisation: Interval variables

Consider $Ext(d) = \{ s \in S : Y_j(s) \subseteq d_j \}$

$$d_j^{(1)} = [l_1, u_1] \quad ; \quad d_j^{(2)} = [l_2, u_2]$$

$$d_j^{(1)} \cup d_j^{(2)} = [\text{Min}\{l_1, l_2\}, \text{Max}\{u_1, u_2\}]$$

Example :

Y_j = time (min) needed to go to work

$$d_j^{(1)} = [5, 15] \quad ; \quad d_j^{(2)} = [10, 20]$$

$$d_j^{(1)} \cup d_j^{(2)} = [5, 20]$$

Generalisation: Multi-valued categorical variables

Consider Ext(d) = { $s \in S : Y_j(s) \subseteq d_j$]

$$d_j^{(1)} = V_1 \quad ; \quad d_j^{(2)} = V_2$$

$$d_j^{(1)} \cup d_j^{(2)} = V_1 \cup V_2$$

Example :

Y_j = jobs of a group of people

$d_j^{(1)}$ = {secretary, teacher} ; $d_j^{(2)}$ = {employee}

$d_j^{(1)} \cup d_j^{(2)}$ = {secretary, teacher, employee}

Generalisation:
Distribution-valued variables

Two possibilities proposed:

→ take for each category the Maximum of its frequencies

→ take for each category the Minimum of its frequencies

Distribution-valued variables: Generalisation by the Maximum

$$d_j^{(1)} \cup d_j^{(2)} = (\, c_{j1}(p_{j1}^{(1)}), \ldots, c_{jk_j}(p_{k_j}^{(1)}) \,) \cup (\, c_{j1}(p_{j1}^{(2)}), \ldots, c_{jk_j}(p_{jk_j}^{(2)}) \,) =$$

$$= (c_{j1}(t_{j1}), \ldots, c_{jk_j}(t_{k_j}) \,) \quad \text{with} \quad t_{j\ell} = \text{Max}\ \{p_{j\ell}^{(1)}, p_{j\ell}^{(2)}\}$$

Example :

Y_j = Type of job

(administration (0.3), teaching (0.7) , secretary (0.0)) ∪

(administration (0.2), teaching (0.6) , secretary (0.2))

= (administration (0.3), teaching (0.7) , secretary (0.2))

Extent: $\{s_i \in S : p_{j\ell}^{(i)} \leq t_{j\ell} \ , \ell = 1, \ldots, k_j\}$

"at most" principle

Distribution-valued variables: Generalisation by the Minimum

$$d_j^{(1)} \cup d_j^{(2)} = (c_{j1}(p_{j1}^{(1)}), \ldots, c_{jk_j}(p_{k_j}^{(1)})) \cup (c_{j1}(p_{j1}^{(2)}), \ldots, c_{jk_j}(p_{jk_j}^{(2)})) =$$

$$= (c_{j1}(r_{j1}), \ldots, c_{jk_j}(r_{k_j})) \qquad \text{with} \qquad r_j = \text{Min}\ \{p_{j\ell}^{(1)}, p_{j\ell}^{(2)}\}$$

Example :

Y_j = Type of job

(administration (0.3), teaching (0.7) , secretary (0.0)) $\cup$

(administration (0.2), teaching (0.6) , secretary (0.2))

= (administration (0.2), teaching (0.6) , secretary (0.0))

Extent: $\{ s_i \in S : p_{j\ell}^{(i)} \geq r_{j\ell} \ , \ell = 1, \ldots, k_j \}$

"at least" principle

Symbolic clustering: the algorithm

Starting with the one-object clusters $\{s_i\}$, i = 1,…,n

At each step, form a cluster p union of p_1, p_2, represented by d such that

- p_1, p_2 can be merged together
- d is more general than d_1, d_2 : $d = d_1 \cup d_2$
- Int (p) = d
- Ext_E (d) = p

Non - uniqueness $\Rightarrow$ numerical criterion

→ Clusters with more specific descriptions are formed first

Symbolic clustering: Generality degree

$d = (d_1, \ldots, d_p)$ $\quad O_j$ bounded

$$G(d) = \prod_{j=1}^{p} G(d_j)$$

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

The more possible members of the extent of d ,
the greater the generality degree of d

Generality degree: Interval-valued variables

$$G(d_j) = \frac{m(V_j)}{m(O_j)} \qquad m(V_j) = \max V_j - \min V_j \quad \text{(range)}$$

Example :

Describing groups of people by age and salary

Age ranges from 15 to 60 , salary ranges from 0 to 10000

Consider a group described by

d= ([20 , 45] , [1000 , 3000]]) = (d_1, d_2)

$$G(d_1) = \frac{45-20}{60-15} = \frac{25}{45} = 0{,}55 \qquad G(d_2) = \frac{3000-1000}{10000-0} = \frac{2000}{10000} = 0{,}2$$

$$G(d) = 0{,}55 \times 0{,}2 = 0{,}11$$

Generality degree: Multi-valued variables

$$G(d_j) = \frac{m(V_j)}{m(O_j)}$$

$m(V_j) = \# V_j$ (cardinal)

Example:

Describing groups of people from the UE, defined on variables gender and nationality (28)

Consider one group described by :

d= ({ M } , {French, English}) = (d_1, d_2)

$$G(d_1) = \frac{1}{2} = 0{,}5$$

$$G(d_2) = \frac{2}{28} = 0{,}07$$

$$G(d) = 0{,}5 \times 0{,}07 = 0{,}035$$

Generality degree: Distribution-valued variables

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

Generalising by the Maximum:

$$G_1(d_j) = \frac{1}{\sqrt{k_j}} \sum_{\ell=1}^{k_j} \sqrt{p_{j\ell}}$$

which is the affinity coefficient (Matusita, 1951) between $(p_{1\ell},\ldots,p_{k_j})$ and the uniform distribution

$G_1(d)$ is maximum (=1) when $p_{j\ell} = 1/k_j$, i=1,…k_j : <u>uniform</u>

This means that we consider a description the more general the more similar it is to the uniform distribution

Generality degree: Distribution-valued variables

$$d_j = (c_{j1}(p_{j1}), \dots , c_{jk_j}(p_{jk_j}))$$

Generalising by the minimum:

$$G_2(d) = \frac{1}{\sqrt{k_j(k_j - 1)}} \sum_{\ell=1}^{k_j} \sqrt{(1 - p_{\ell j})}$$

Again, $G_2(d)$ is maximum (=1)

when $p_{j\ell} = 1/k_j$, i=1,...k : <u>uniform</u>

Symbolic clustering: the algorithm

Starting with the one-object clusters $\{ s_i \}, i = 1,\ldots,n$

At each step, form a cluster p union of p_1 , p_2 , represented by d such that

- p_1, p_2 can be merged together
- d is more general than d_1, d_2 : $d = d_1 \cup d_2$
- $Int(p) = d$ and $Ext_E(d) = p$: (p, d) is a concept
- $G(d)$ is minimum

Symbolic clustering: the algorithm

The algorithm builds a hierarchy / pyramid on S :
each cluster is associated to a description
whose extent is the cluster itself

CLUSTER $\leftrightarrow$ CONCEPT

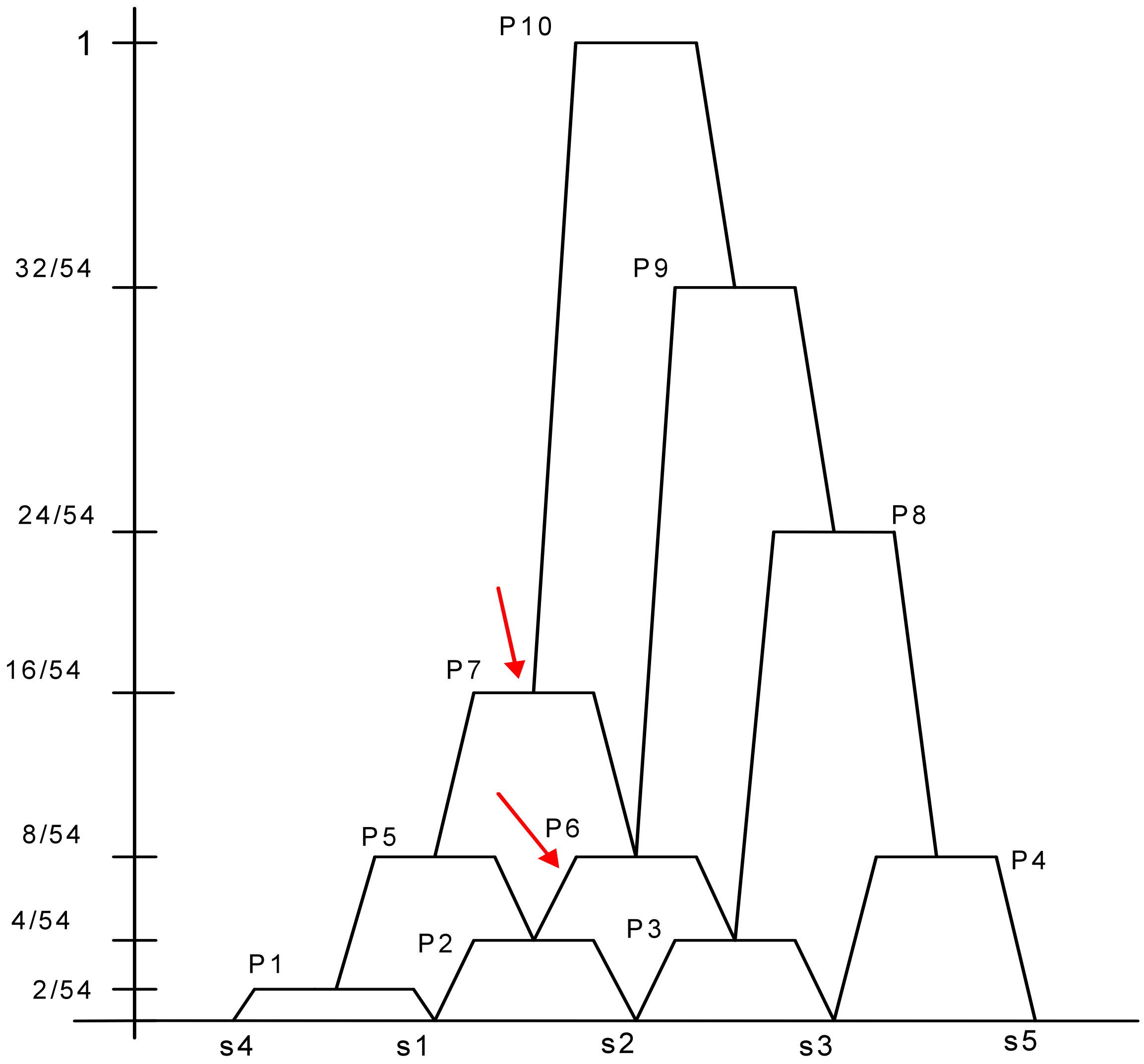
CLUSTER = (p, d) p = Ext d, d = Int(p)

$\longrightarrow$ automatic representation of the clusters

Example

	Y_1	Y_2	Y_3	Y_4
s_1	1	1	1	2
s_2	1	2	1	3
s_3	1	2	2	2
s_4	2	1	1	2
s_5	3	3	2	1

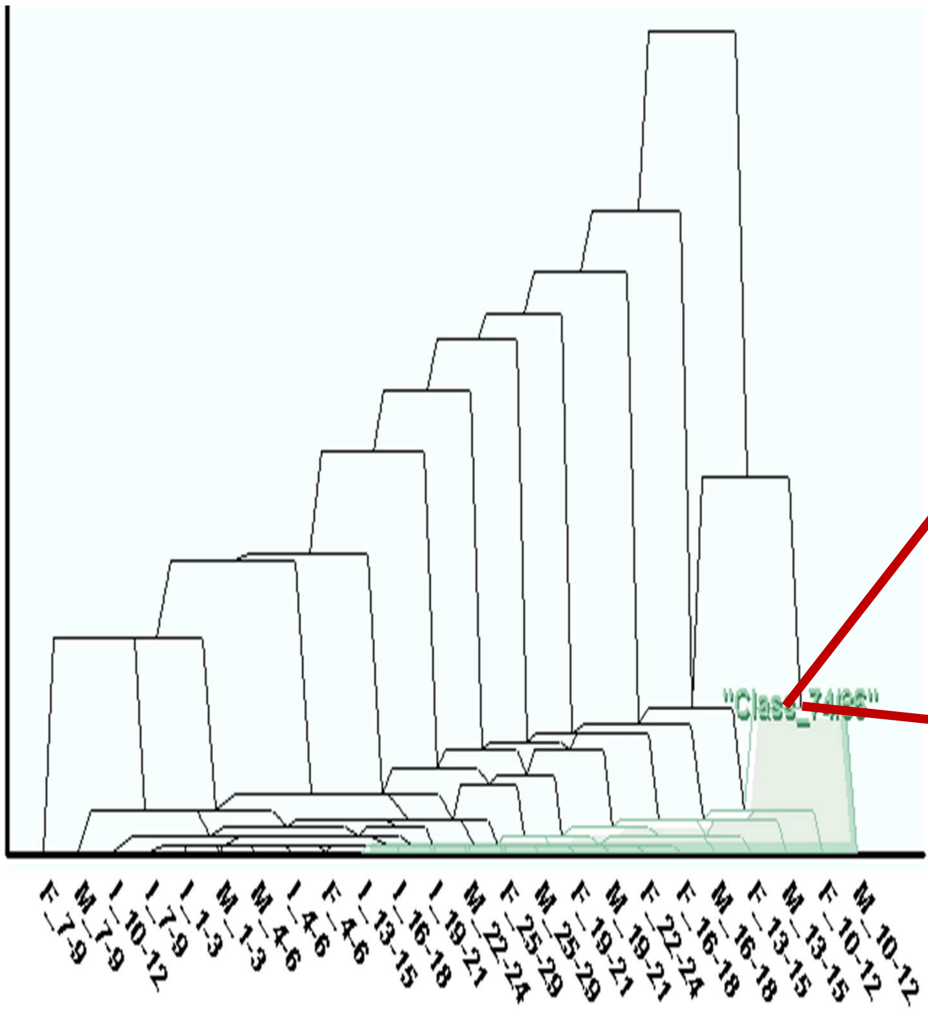
Y_j : Numerical multi-valued variables

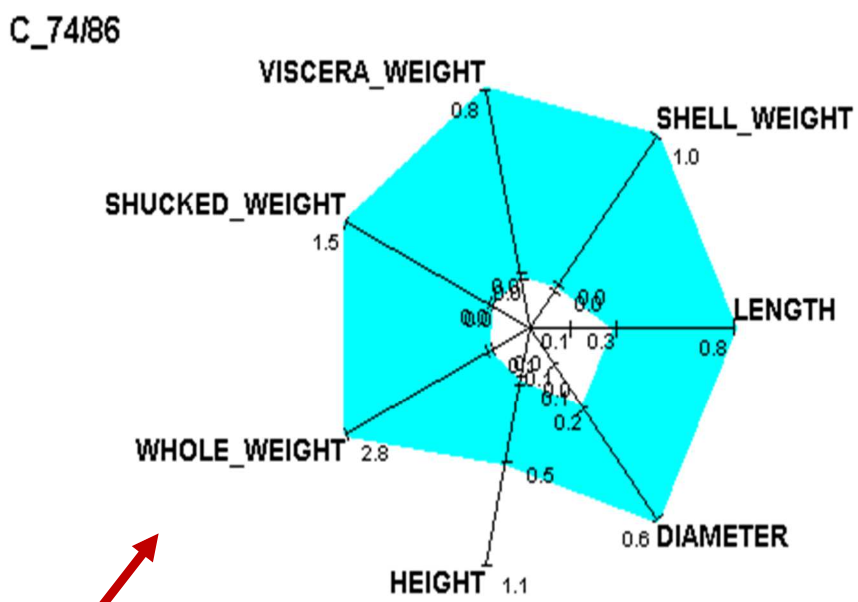


P_6 : ($\{s_1, s_2, s_3\}$; ({1} ,{1,2} , {1,2}, {2,3}))

P_7 : ($\{s_1, s_2, s_3, s_4\}$; ({1,2} ,{1,2} , {1,2}, {2,3}))

Abalone data
Symbolic pyramid





name	"Class_74/86"	
label	"C_74/86"	
height	0.176758	
symbolic object description	variable list (conjunction of)	
	name	value
	AB00	[0.29, 0.815]
	AC00	[0.225, 0.65]
	AD00	[0.06, 0.515]
	AE00	[0.12, 2.8255]
	AF00	[0.0415, 1.488]
	AG00	[0.026, 0.76]
	AH00	[0.04, 1.005]
base object list	I_13-15, I_16-18, I_19-21, M_22-24, F_25-29, M_25-29, F_19-21, M_19-21, F_22-24, F_16-18, M_16-18, F_13-15, M_13-15, F_10-12, M_10-12	

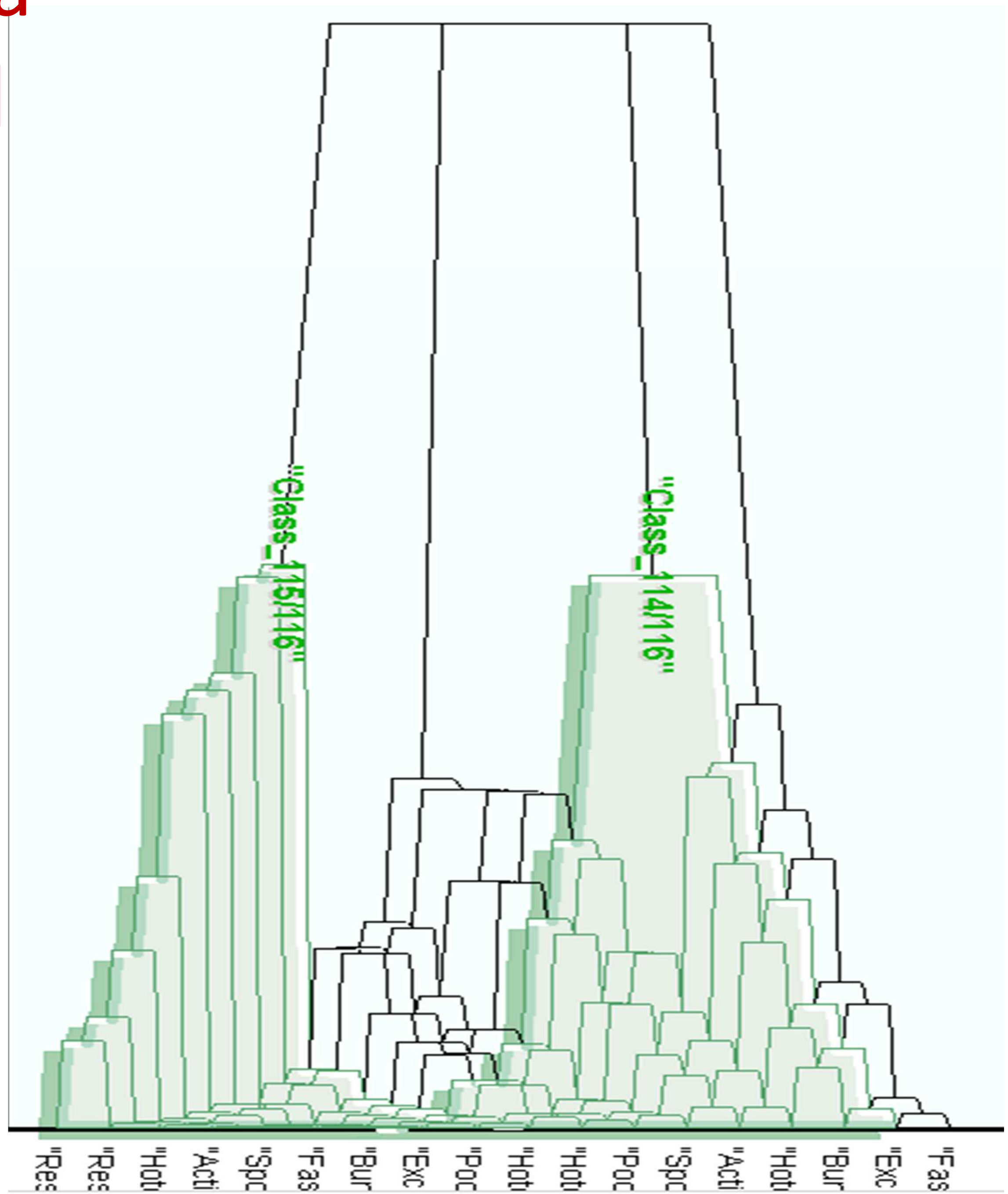
Symbolic pyramid : Cluster description

name	"Class_74/86"	
label	"C_74/86"	
height	0.176758	
symbolic object description	variable list (conjunction of)	
	name	value
	AB00	[0.29, 0.815]
	AC00	[0.225, 0.65]
	AD00	[0.06, 0.515]
	AE00	[0.12, 2.8255]
	AF00	[0.0415, 1.488]
	AG00	[0.026, 0.76]
	AH00	[0.04, 1.005]
base object list	I_13-15, I_16-18, I_19-21, M_22-24, F_25-29, M_25-29, F_19-21, M_19-21, F_22-24, F_16-18, M_16-18, F_13-15, M_13-15, F_10-12, M_10-12	

Travel agency data

	pays_client	resort	intervallePrice	age_range	pays
Restaurant in U	US (0.45), Germa (0.09), Japan (0.45)	Baham (0.64), Hawai (0.36)	[95.00 : 150.00]	25-39 (0.35), 51-70 (0.27), 18-24 (0.38)	US
Hotel Room in U	US (0.33), Germa (0.33), Japan (0.33)	Baham (0.50), Hawai (0.50)	[192.00 : 195.00]	25-39 (0.32), 18-24 (0.68)	US
Hotel Room in F	US (0.33), Germa (0.33), Japan (0.33)	Frenc (1.00)	[170.00 : 170.00]	25-39 (0.33), 18-24 (0.67)	Franc
Restaurant in F	US (0.50), Japan (0.50)	Frenc (1.00)	[85.00 : 85.00]	25-39 (0.50), 18-24 (0.50)	Franc
Excursion in US	US (0.50), Japan (0.50)	Baham (0.50), Hawai (0.50)	[100.00 : 100.00]	25-39 (0.04), 40-50 (0.96)	US
Bungalow in US	US (0.33), Germa (0.33), Japan (0.33)	Baham (0.50), Hawai (0.50)	[150.00 : 160.00]	25-39 (0.04), 40-50 (0.96)	US
Excursion in Fr	US (0.50), Japan (0.50)	Frenc (1.00)	[175.00 : 175.00]	40-50 (1.00)	Franc
Bungalow in Fra	US (0.33), Germa (0.33), Japan (0.33)	Frenc (1.00)	[120.00 : 120.00]	40-50 (1.00)	Franc
Hotel Suite in	US (0.33), Germa (0.33), Japan (0.33)	Baham (0.50), Hawai (0.50)	[292.00 : 295.00]	51-70 (0.96), Over (0.04)	US
Poolside Bar in	US (0.50), Japan (0.50)	Baham (0.50), Hawai (0.50)	[80.00 : 85.00]	51-70 (0.96), Over (0.04)	US
Hotel Suite in	US (0.33), Germa (0.33), Japan (0.33)	Frenc (1.00)	[270.00 : 270.00]	51-70 (1.00)	Franc
Poolside Bar in	US (0.50), Japan (0.50)	Frenc (1.00)	[120.00 : 120.00]	51-70 (1.00)	Franc
Activities in U	Germa (1.00)	Baham (0.50), Hawai (0.50)	[150.00 : 200.00]	18-24 (1.00)	US
Activities in F	Germa (1.00)	Frenc (1.00)	[50.00 : 50.00]	18-24 (1.00)	Franc
Sports in US	Germa (1.00)	Baham (0.50), Hawai (0.50)	[100.00 : 150.00]	51-70 (0.96), Over (0.04)	US
Sports in Franc	Germa (1.00)	Frenc (1.00)	[190.00 : 190.00]	51-70 (1.00)	Franc
Fast Food in US	Germa (1.00)	Baham (0.50), Hawai (0.50)	[80.00 : 105.00]	25-39 (0.04), 40-50 (0.96)	US
Fast Food in Fr	Germa (1.00)	Frenc (1.00)	[90.00 : 90.00]	40-50 (1.00)	Franc

Travel agency data
Symbolic pyramid



name	"Class_115/116"	
label	"C_115/116"	
height	0.0441423	
symbolic object description	variable list (conjunction of)	
	name	value
	AB00	(janvier(0.25), février(0.25), mars(0.25), avril(0.25), mai(0.25), juin(0.25), juillet(0.25), août(0.25), s...
	AC00	[18, 68]
	AD00	[4, 12]
	AE00	[2, 8]
	AF00	(South(0.5), West(0.5), East Coast(0.0151515), Mid West(0.5), Bavaria(1), East Germany(1), East ...
	AG00	(US(0.5), Germany(1), Japan(0.5))
	AH00	(Bahamas Beach(0.636364), French Riviera(1), Hawaiian Club(0.363636))
base object list	"AA00", "AA03", "AA02", "AA13", "AA15", "AA17", "AA07", "AA06", "AA11", "AA10"	

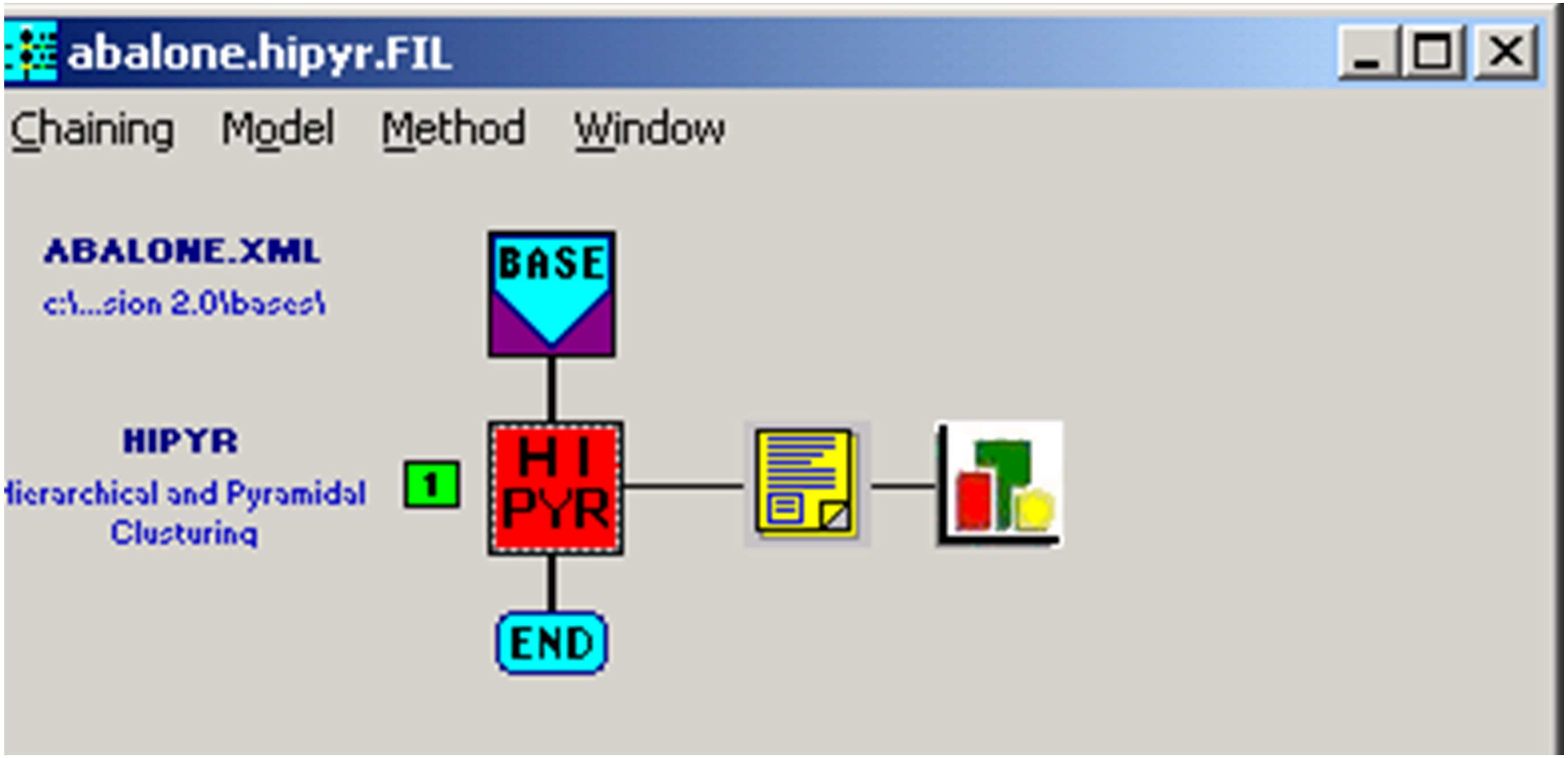
The *HIPYR* module of the *SODAS* software

Objective :

Perform Hierarchical or Pyramidal clustering on a symbolic data set

- from a dissimilarity matrix
  $\rightarrow$ numerical clustering
- directly based on the data set
  $\rightarrow$ symbolic clustering: clusters are concepts

The *HIPYR* module of the *SODAS* software



HIPYR : Main Parameters

Structure: Hierarchy or Pyramid

Data Source:

- Dissimilarity Matrix (Numerical Clustering)
- Symbolic objects (Symbolic Clustering)

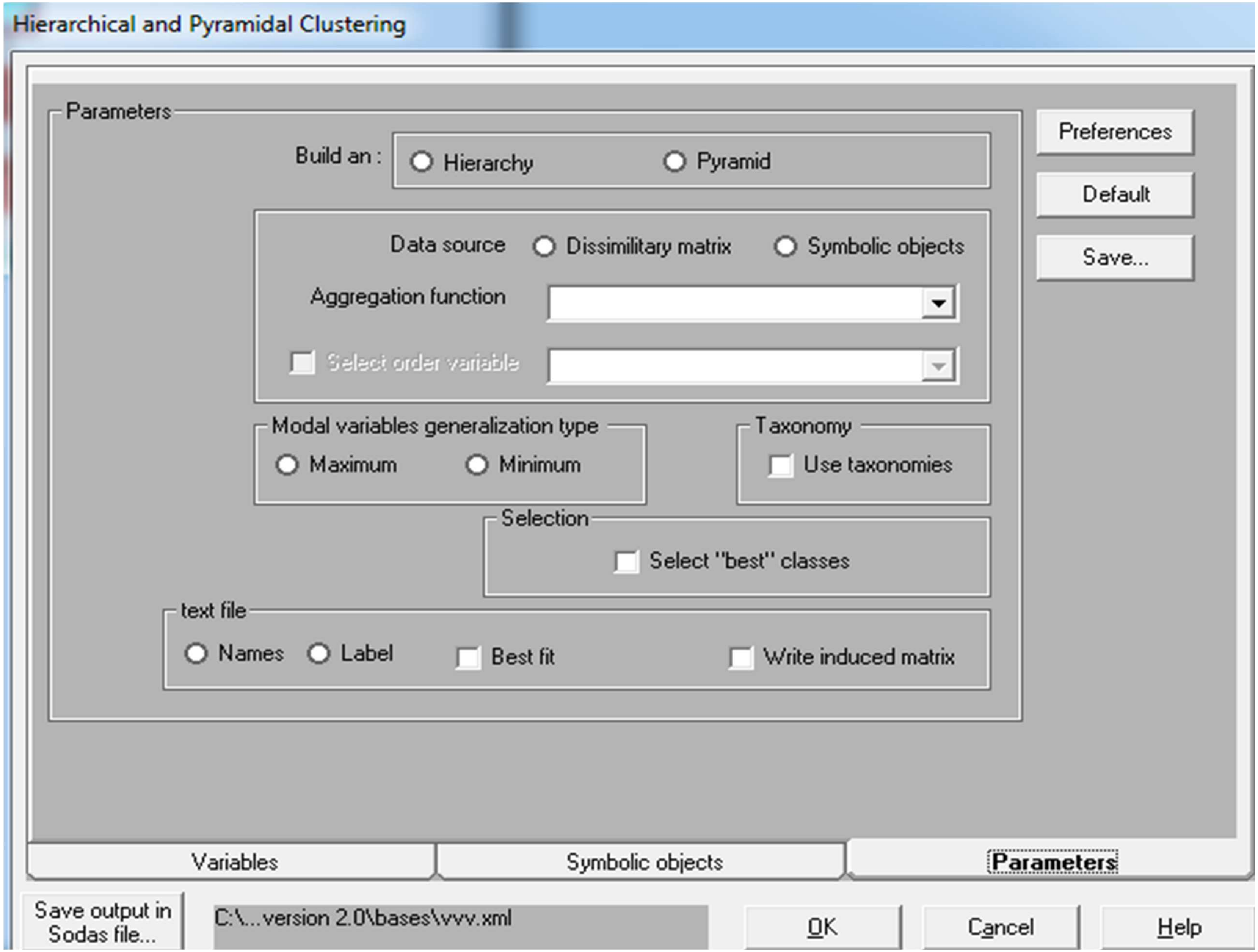
Aggregation Index:

- Numerical Clustering: Maximum, Minimum, Average, Diameter
- Symbolic Clustering: Minimum Generality
  Minimum Increase in Generality

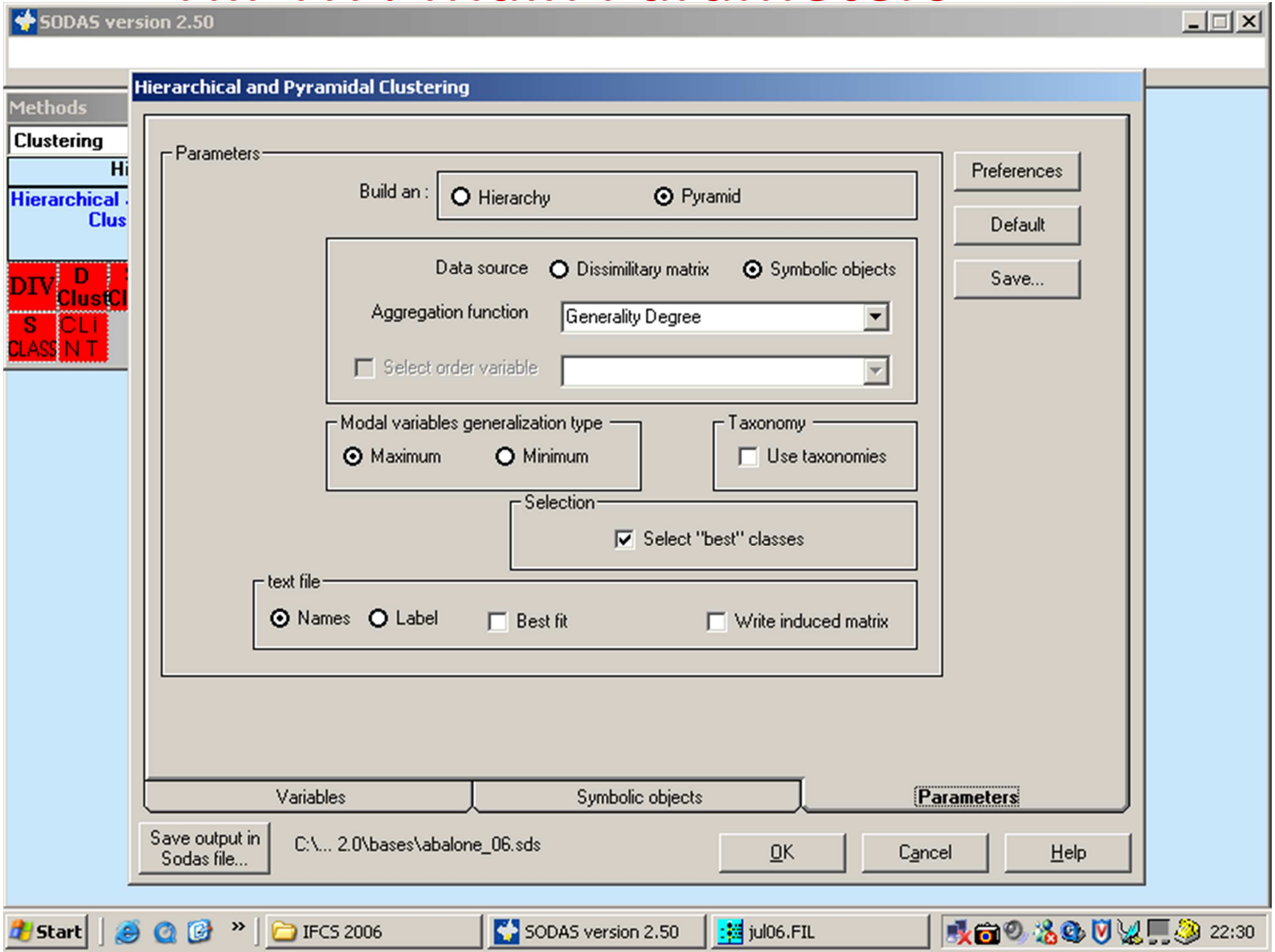
HIPYR : Main Parameters

- Order Variable (optional) : quantitative single variable; to impose an order compatible with the pyramid
- Modal variables generalization :
  - Maximum
  - Minimum
- Use Taxonomies for generalization (nominal or categorical multi-valued variables) : Y, N
- Select "best" classes : Y, N
- Write induced dissimilarity/generality matrix : Y, N

HIPYR : Main Parameters



HIPYR : Main Parameters



Induced dissimilarity/generality matrix

For each pair of elements of S , s_i , $s_{i'}$

$d^*(s_i, s_{i'})$ = index (height) of the "smallest" class that contains s_i and $s_{i'}$

$$d^*(s_i, s_{i'}) = \text{Min } \{f(C), s_i \in C, s_{i'} \in C\}$$

Evaluation of the obtained indexed hierarchy / pyramid: Comparision between the initial and the induced dissimilarity/generality matrices.

Evaluation value

For s_i $s_{j,}$ i, j, =1,..., n, $d(s_i, s_j)$:

- the given dissimilarity matrix (numerical clustering)
- generality degree of $s_i \cup s_j$ (symbolic clustering)

$$EV = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(d(s_i, s_j) - d^*(s_i, s_j)\right)^2}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d(s_i, s_j)}$$

Cluster selection

Identify the most interesting clusters :

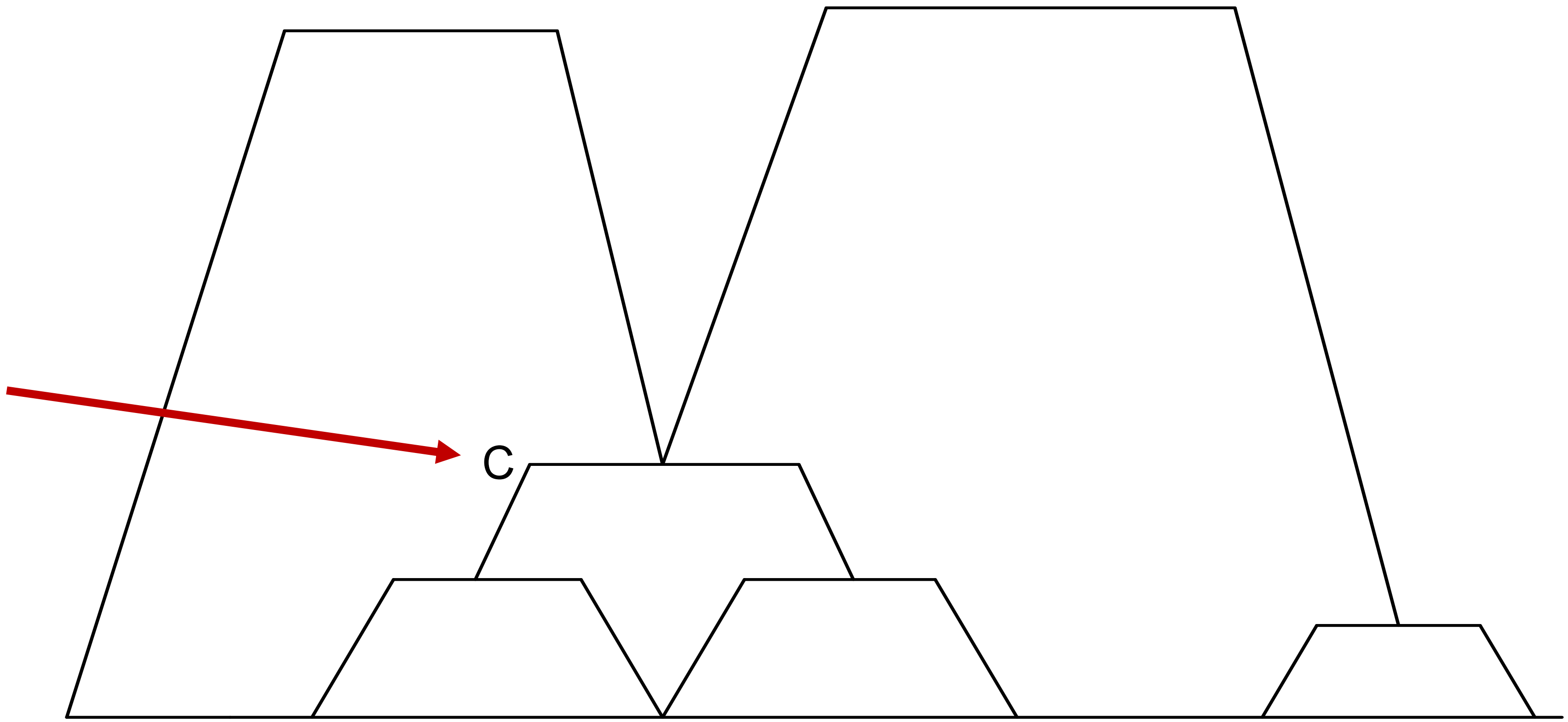
A cluster is "interesting" if its variability is small as compared to its predecessors.

Variability indicated by index values f(h).

Compute mean value and standard deviation of height increase values.

A class is selected if the corresponding increase value is more than 2 stand. dev. over the mean value.

Cluster selection



HIPYR Output

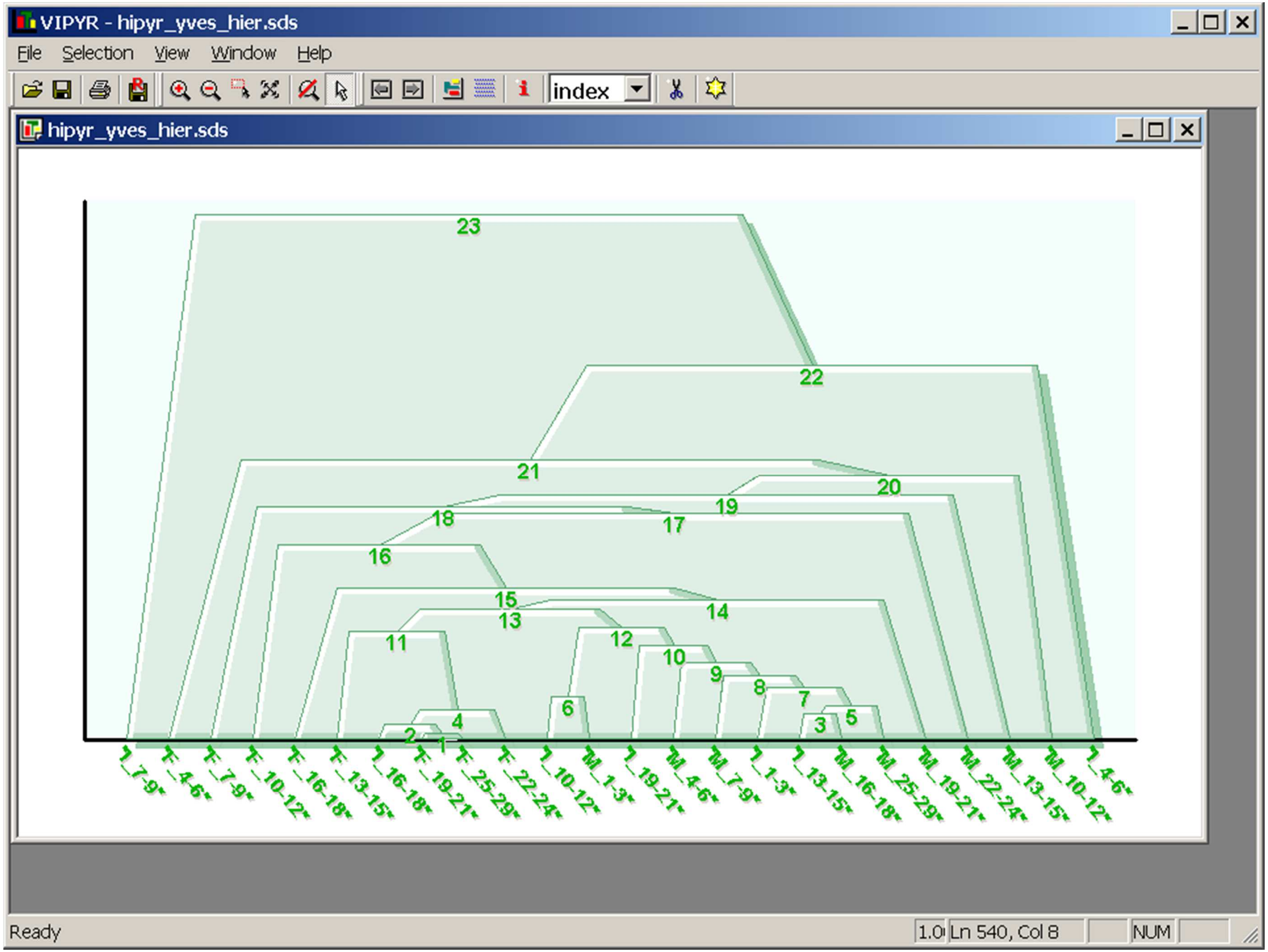
- Text file
- Sodas file
- Interactive Graphical Representation (VPYR)

HIPYR Output

The output listing contains:

- The labels of the individuals
- The labels of the variables
- The description of each node :
  - the symbolic object associated to each node
  - its extent
- Evaluation value
- Selected clusters, if asked for
- The induced matrix, if asked for

Graphical Representation



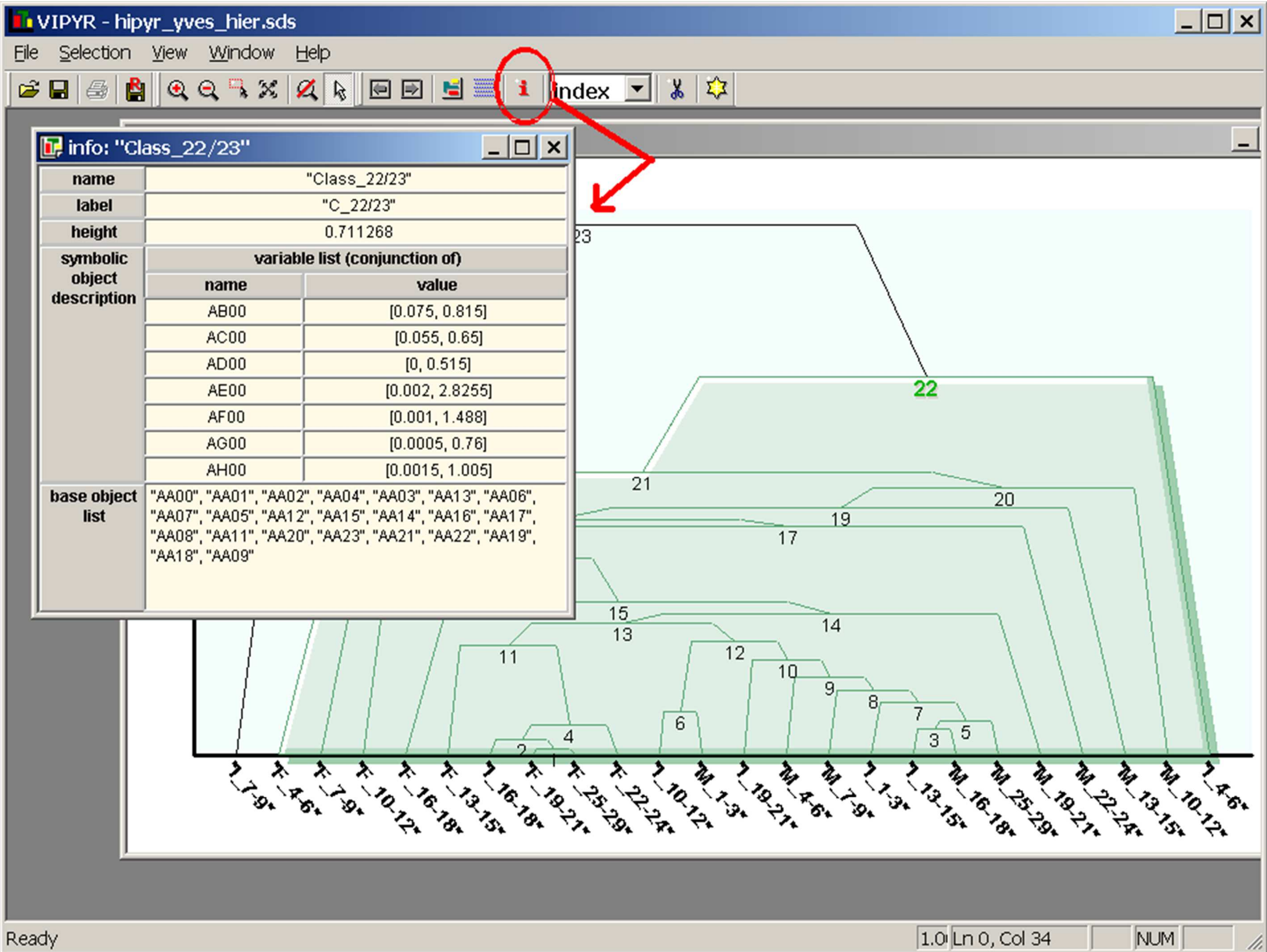
Graphical Representation: Options

A cluster is selected by clicking on it.

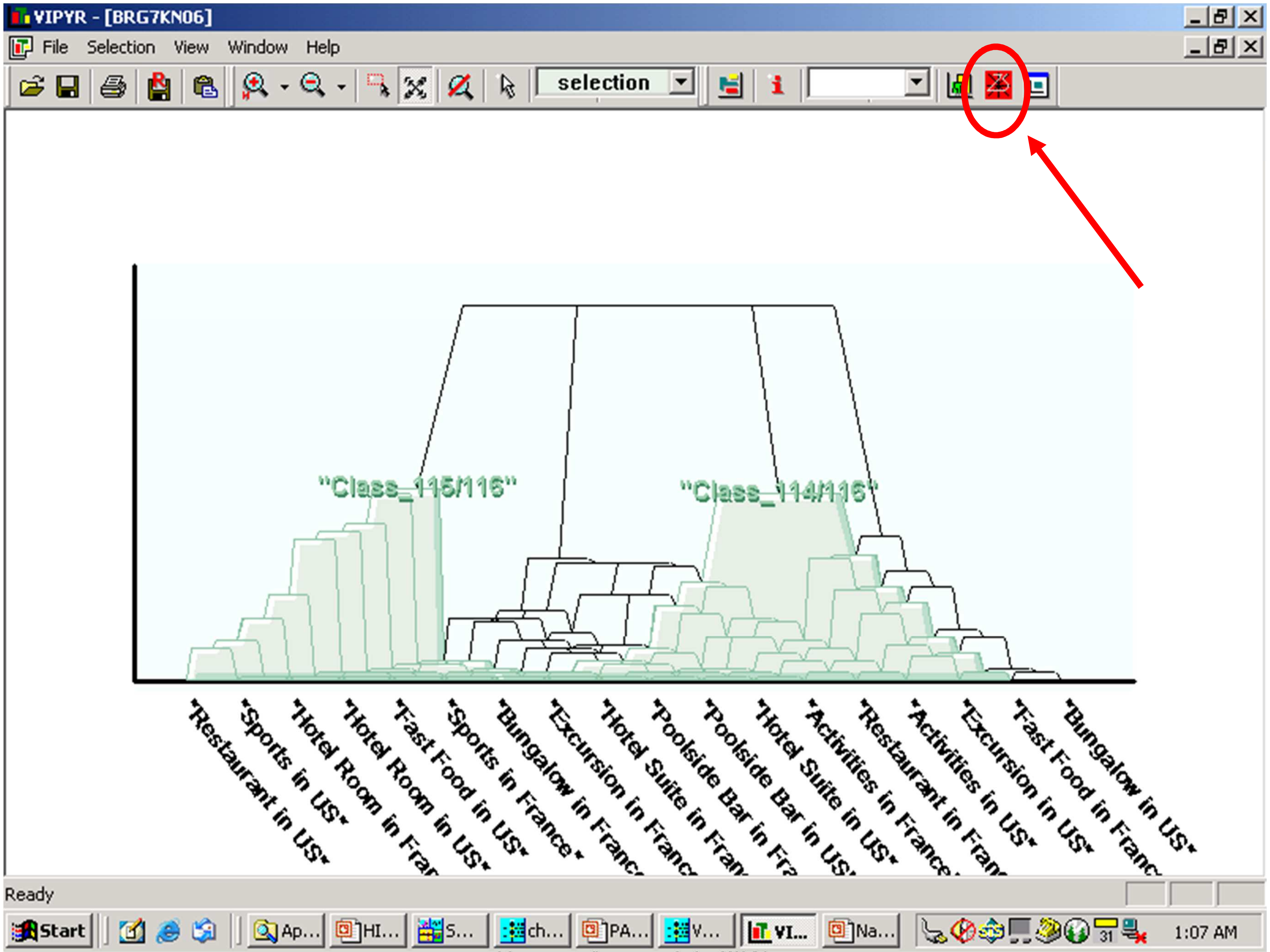
Description of the cluster in terms of

- list of chosen variables
- representation by a Zoom Star

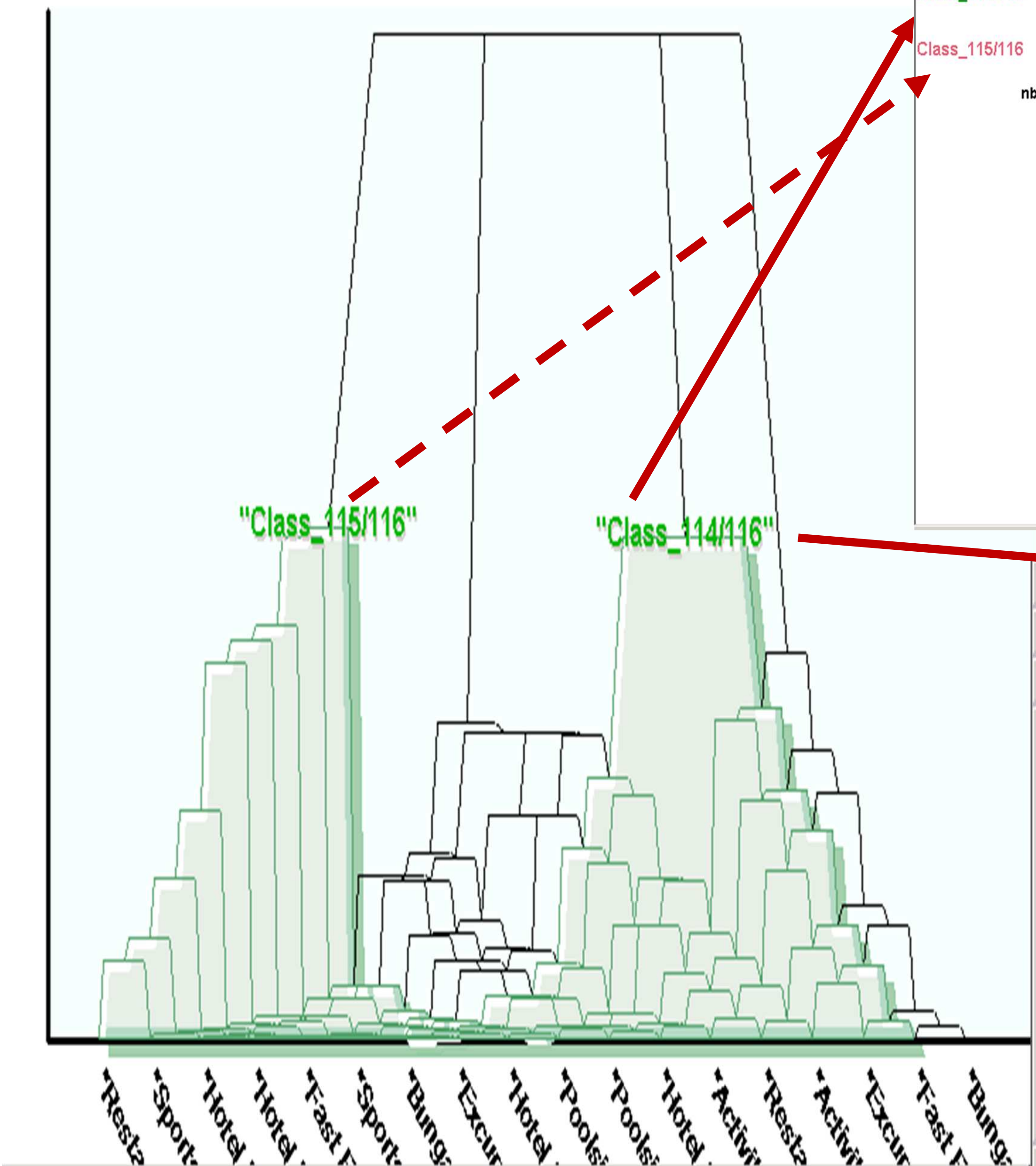
Graphical Representation: Options

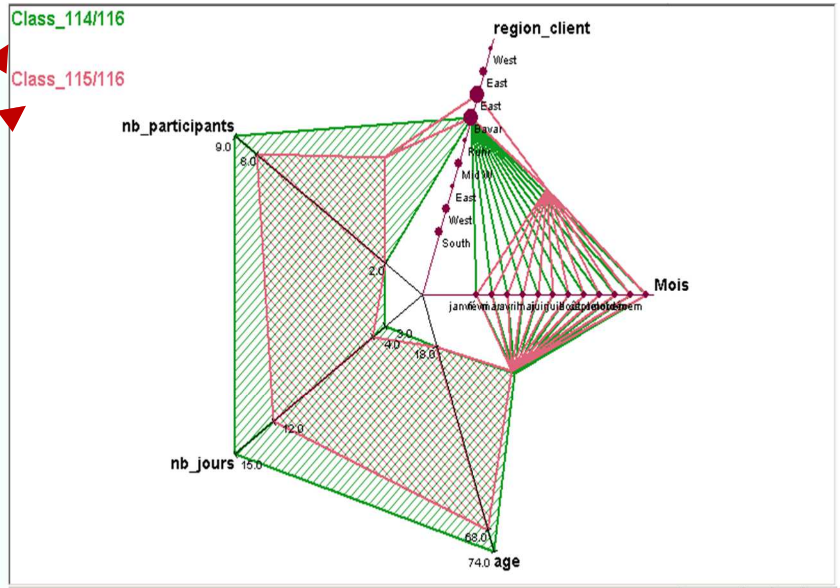


Graphical Representation: Options



HIPYR - VPYR





name	"Class_114/116"	
label	"C_114/116"	
height	0.043303	
symbolic object description	variable list (conjunction of)	
	name	value
	AB00	(janvier(0.25), février(0.25), mars(0.25), avril(0.25), mai(0.25), juin(0.25), juillet(0.25), août(0.25), septembre...
	AC00	[18, 74]
	AD00	[3, 15]
	AE00	[2, 9]
	AF00	(South(0.319444), West(0.5), East Coast(0.0208333), Mid West(0.5), Ruhr(0.0416667), Bavaria(0.958333),...
	AG00	(US(0.5), Germany(1), Japan(0.5))
	AH00	(Bahamas Beach(0.5), French Riviera(1), Hawaiian Club(0.5))
base object list	"AA07", "AA06", "AA11", "AA10", "AA08", "AA09", "AA14", "AA12", "AA01", "AA05", "AA04"	

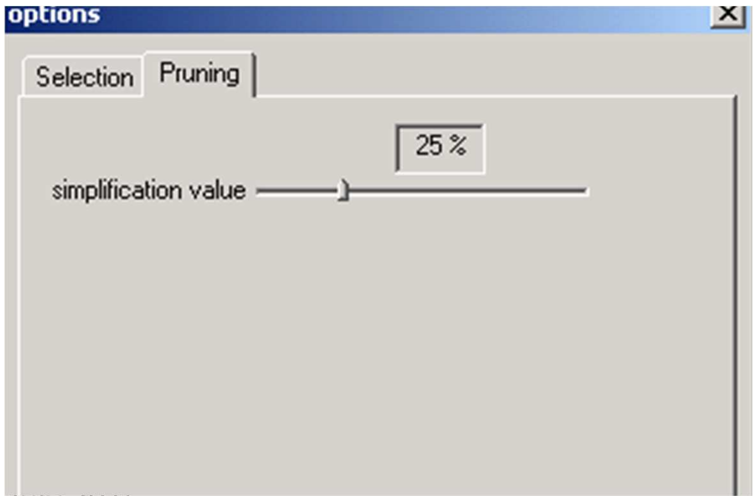
Graphical Representation: Options

Pruning the hierarchy or pyramid using the aggregation heights as a criterion.

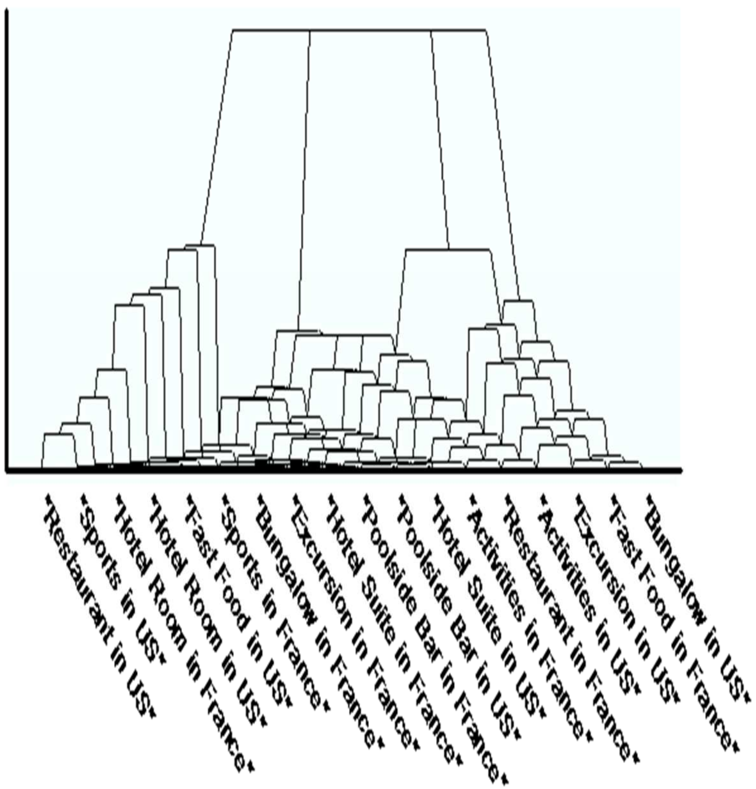
Suppressing cluster p if :

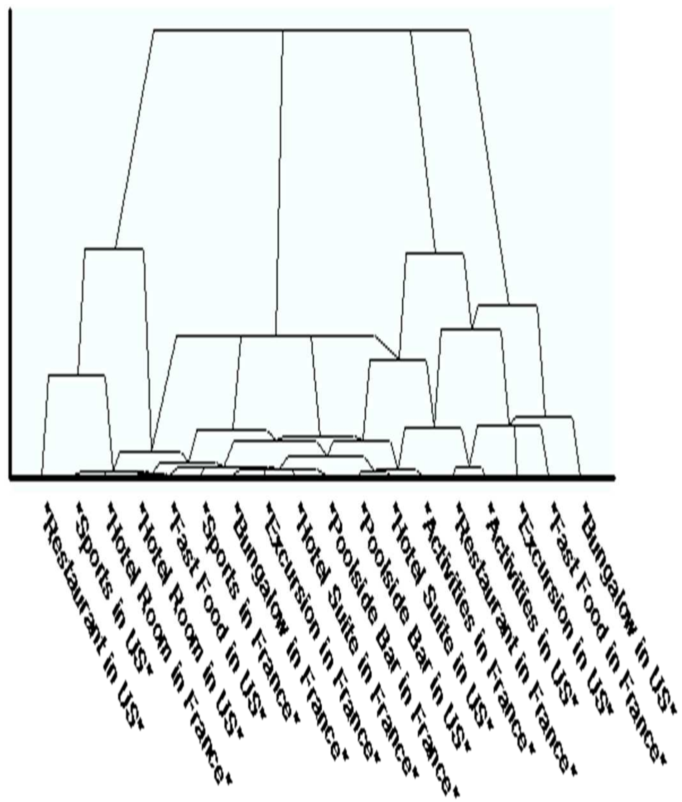
$f(p') - f(p) < \alpha f(S) \wedge$ p has a single predecessor

Rate of simplification α chosen by the user, new graphic window with the simplified structure.



Graphical Representation: Pruning



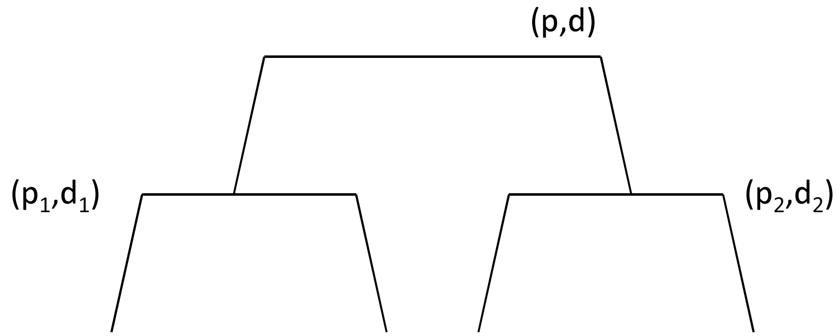


Rule Generation

Hierarchy/pyramid built from a symbolic data table: rules may be generated and saved in a specified file

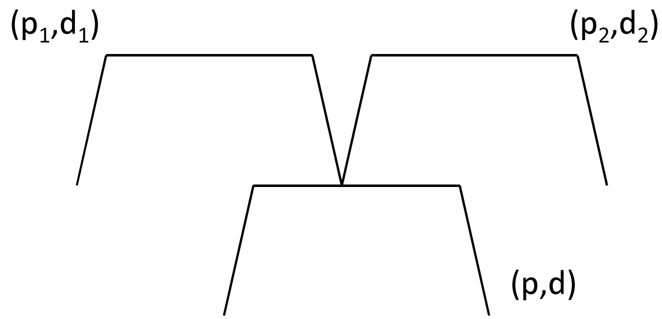
Fission method :

$$d \Rightarrow d_1 \vee d_2$$

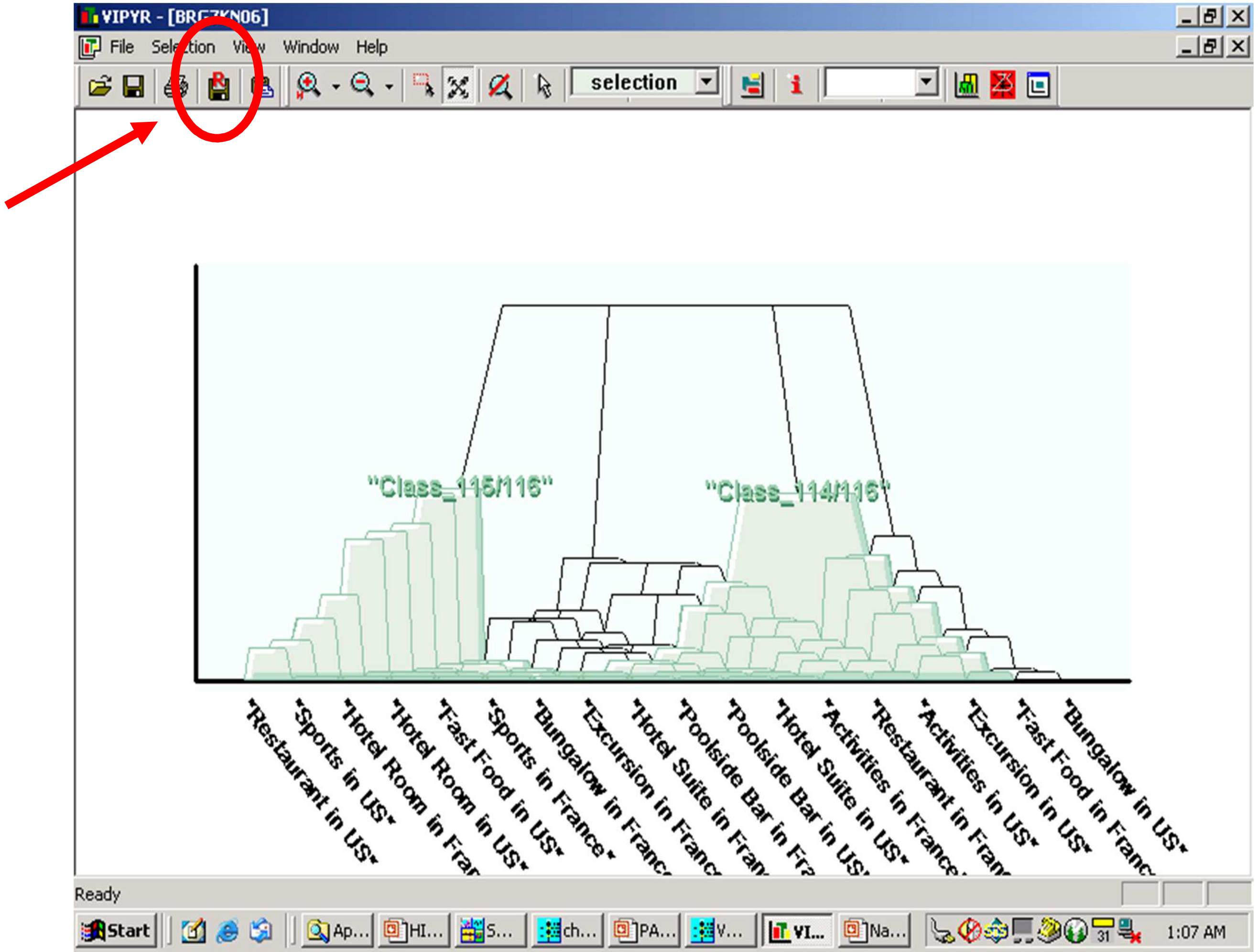


Fussion method (pyramids only) :

$$d_1 \wedge d_2 \Rightarrow d$$



Rule Generation



Graphical Representation: Options

Reduction

Should the user be interested in a particular cluster, he may obtain a window with the structure restricted to this cluster and its successors.

Graphical Representation: Reduction