

Induction Graphs (*I.G.*)

Professor D.A. Zighed

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

What are *I.G.* able to do ?

- Extract knowledge from data
- Build predictive models
- Describe a phenomenon
- Identify patterns
- . . .

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

How do *I.G.* proceed ?

- By searching the best partition on the training set,
- Giving a lattice graph,
- Giving the rules.

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

Examples . . .

■ Cementation placement quality

Why in the same hole can't we always have a good cementation ?

- *Image identification,*
- *Medical diagnosis,*
- *Chemical identification,*
- *Management,*
- . . .

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

I.G.'s Process

- Training

- Validation

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

Training

- Training set : Ω , $\text{Card}(\Omega)=n$

File of data :

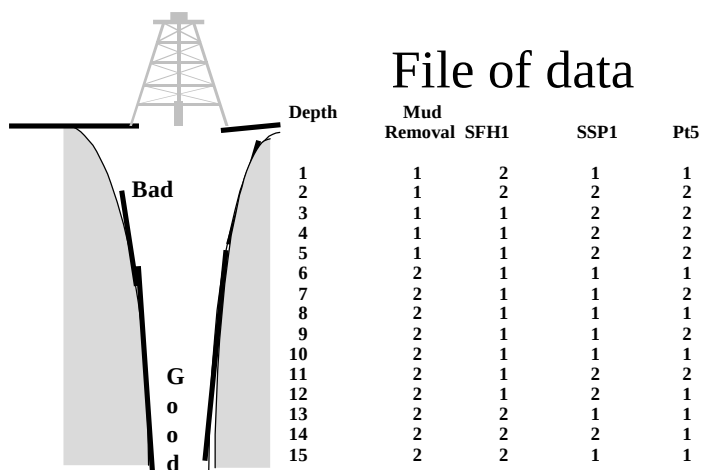
Depths X parameters of fluid placement, of reology, ...

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

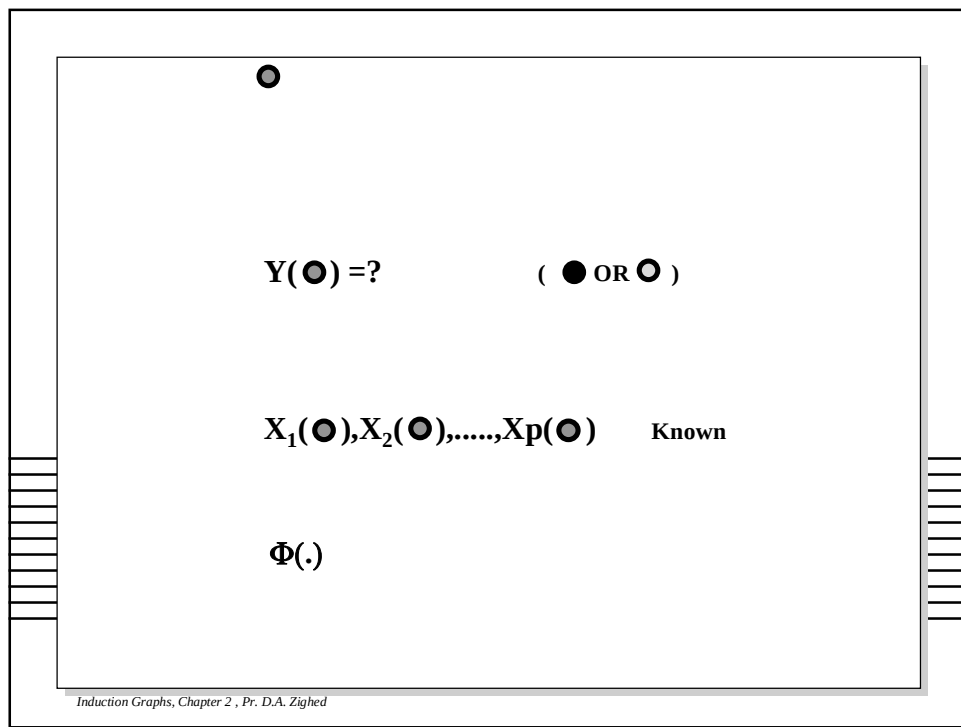
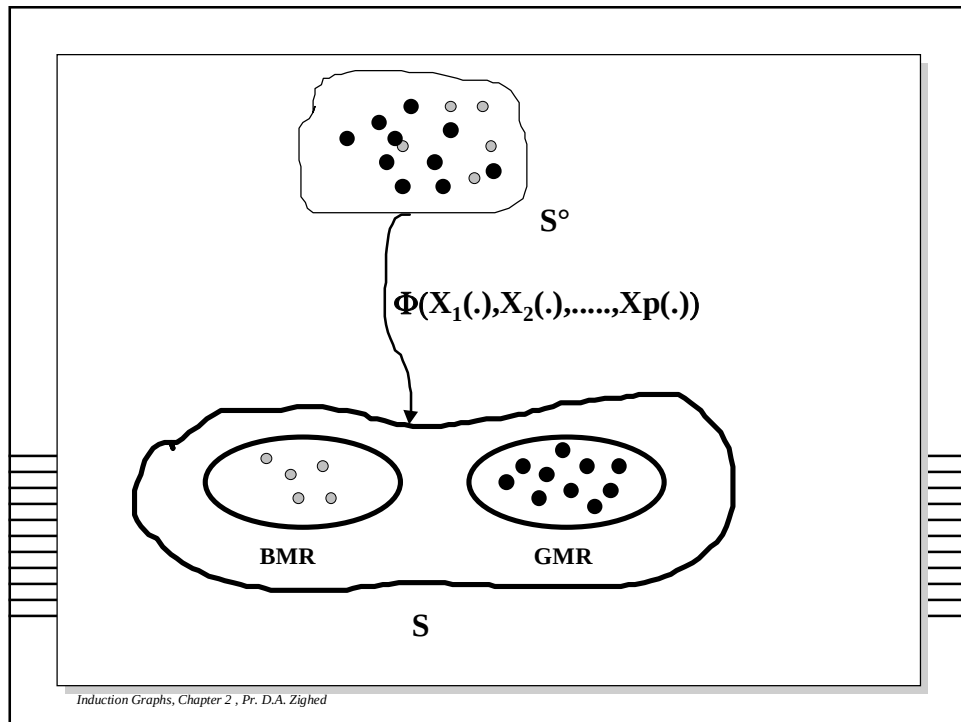
		Parameters					
		Pt1	Pt2	SSP1		Pmvt	Cementation
		$x_1(.)$	$x_2(.)$	$x_3(.)$		$x_p(.)$	$y_p(.)$
Ω							
ω_1		1,34	1,8	10		Yes	Good
ω_2		1,34	1	-3		No	Bad
ω_n		1	2	4		No	Bad

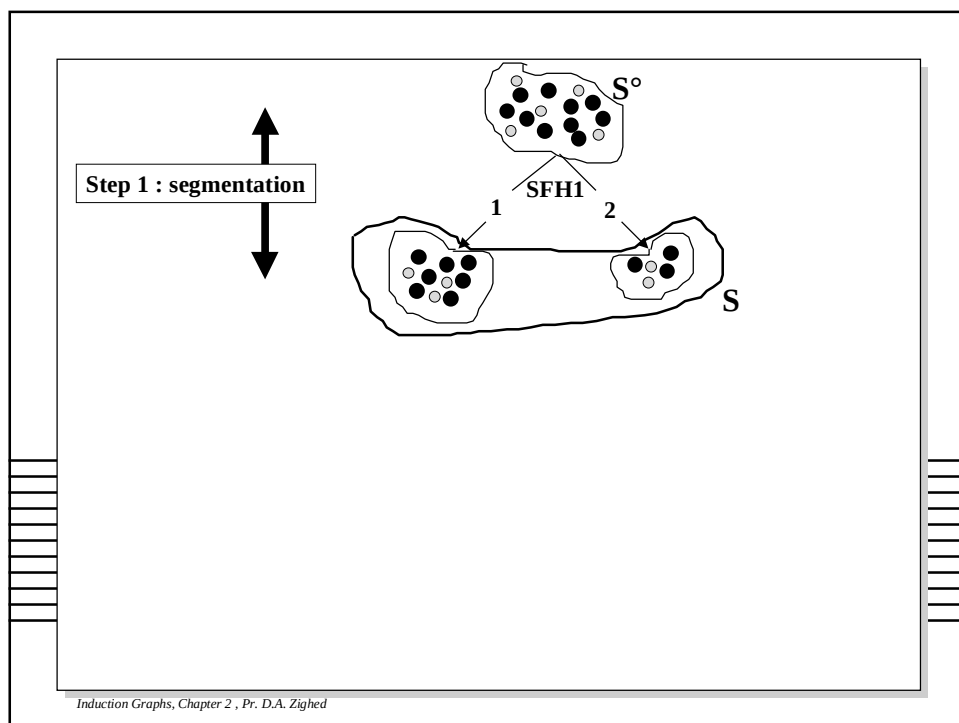
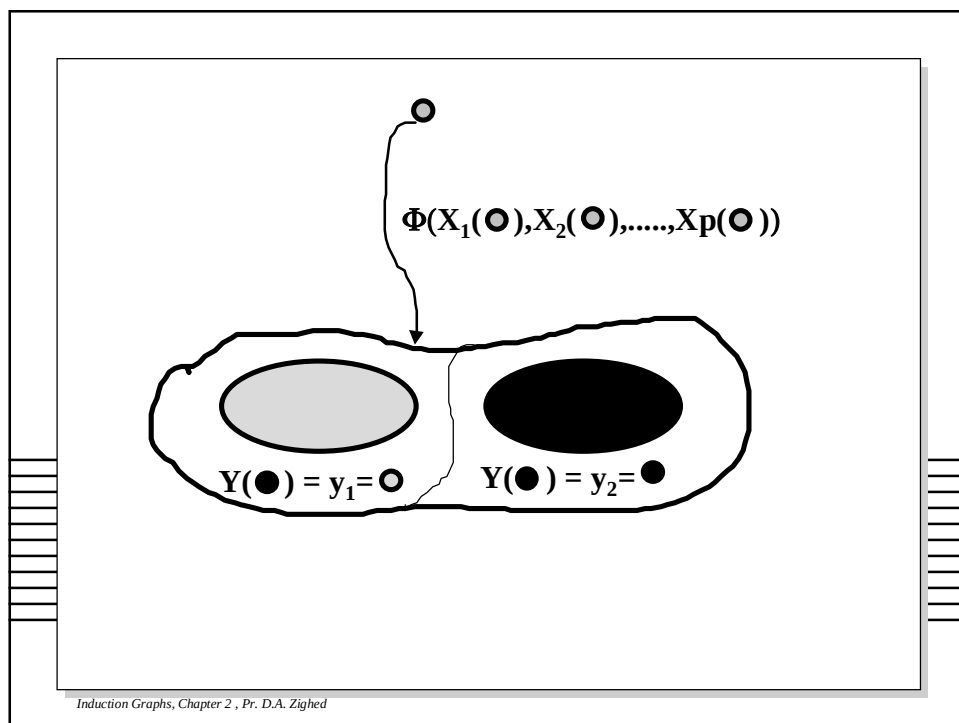
Endogenous variable

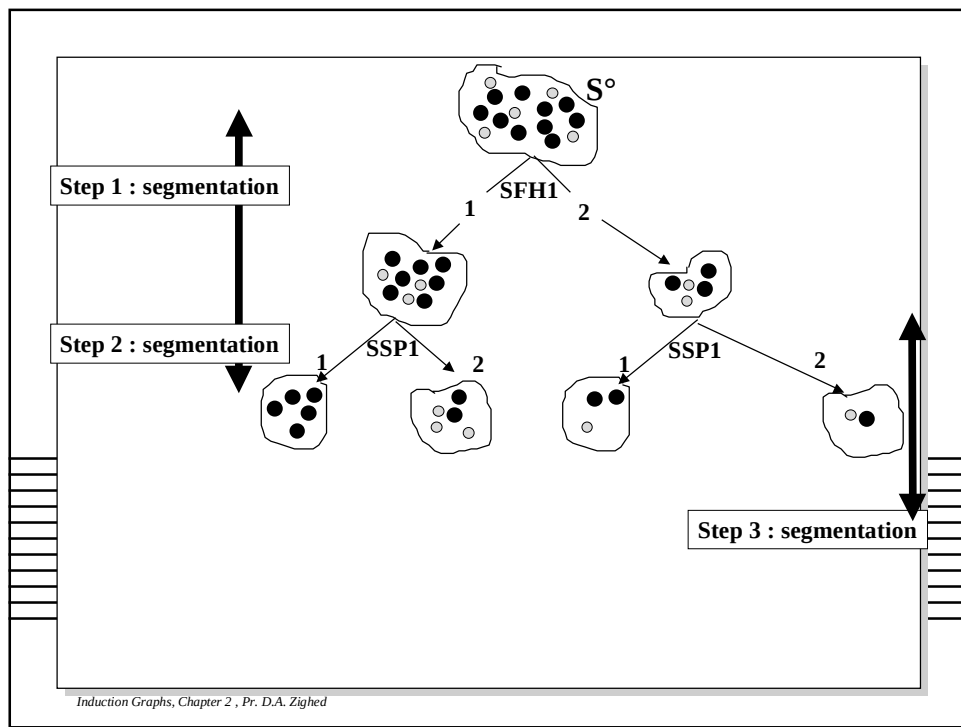
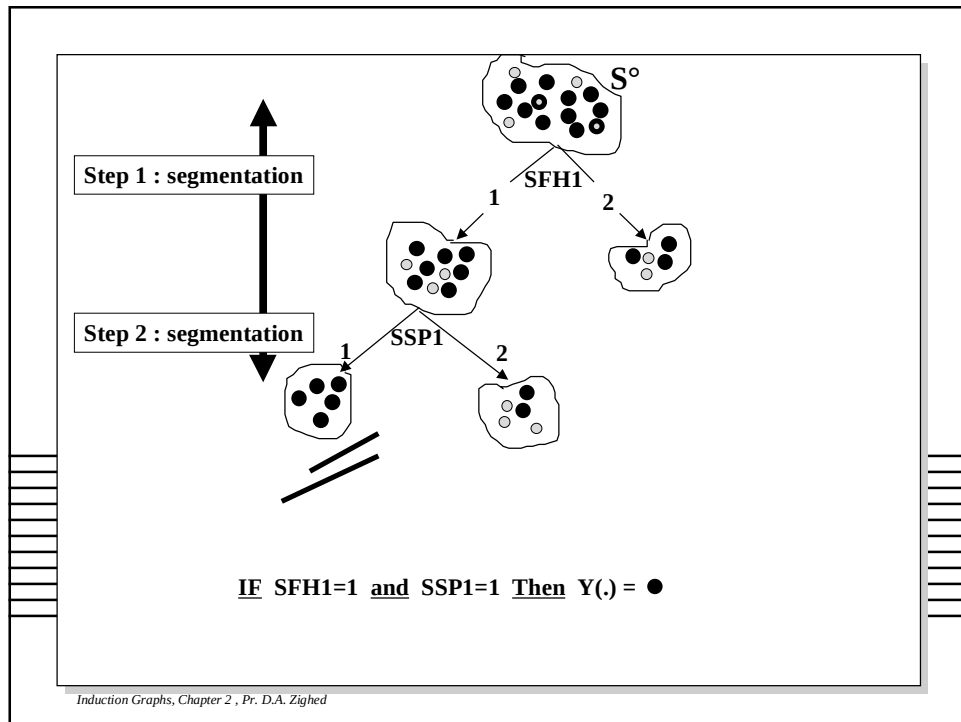
Induction Graphs, Chapter 2 , Pr. D.A. Zighed

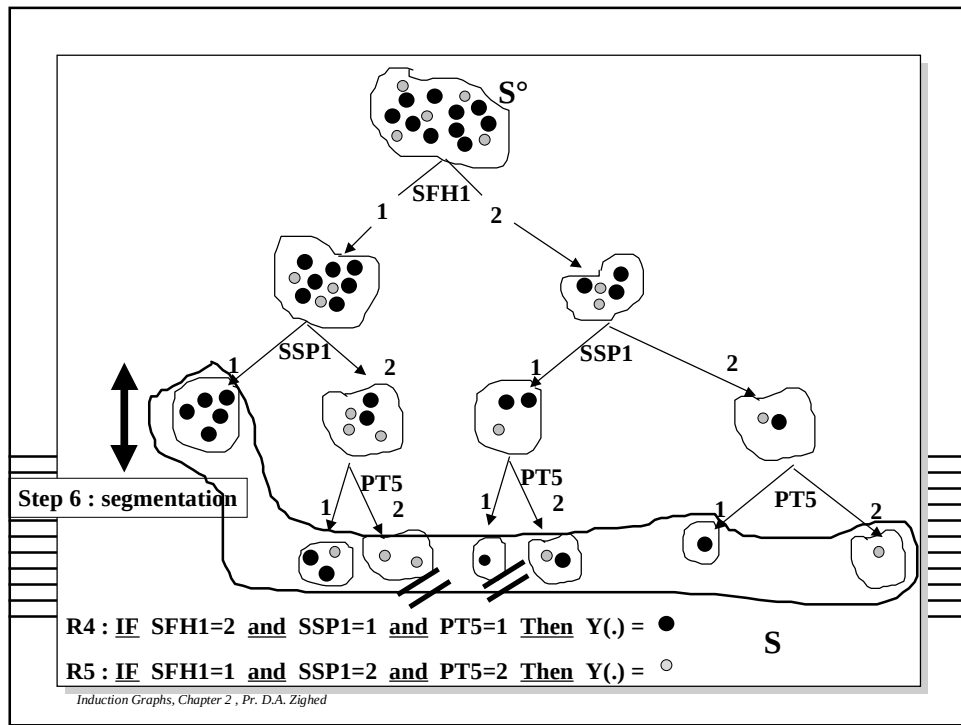
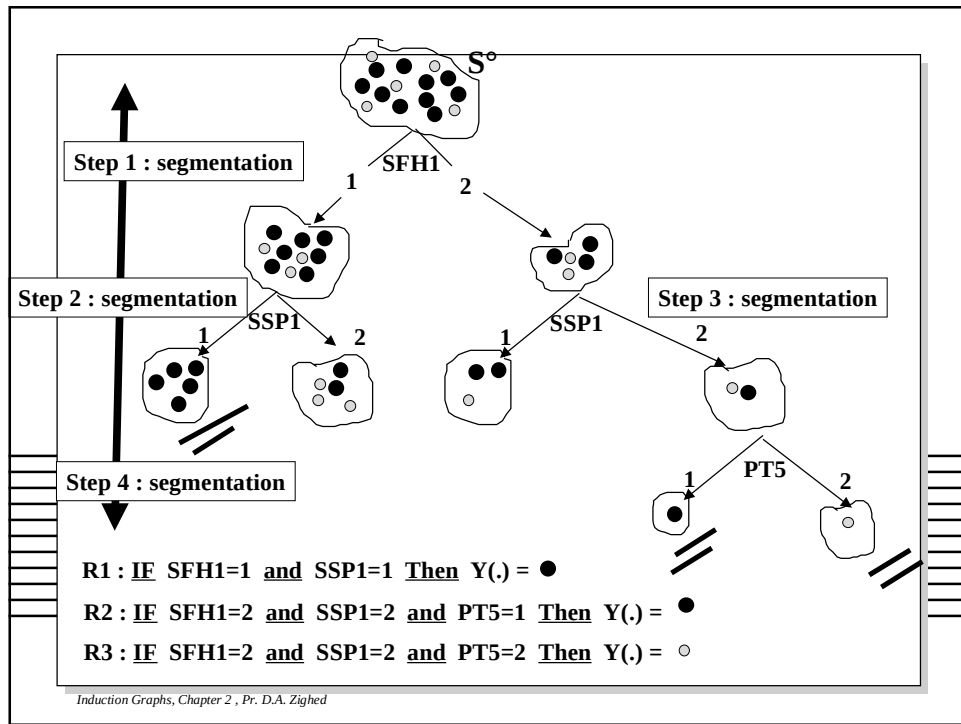


Induction Graphs, Chapter 2 , Pr. D.A. Zighed









Questions

- ✓ **What is the validity of the rules generated by I.G's process ?**
 - How does one classify an anonymous point ?
 - How does one evaluate the rule efficiency ?
 - What is the general validity of the process ?
- ✓ **Is segmentation the best way for obtaining the sequence of partitions ?**

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

	Ciment.	SFH1	SSP1	Pt5
20	?	2	2	1
21	?	2	2	2
32	?	2	2	1
4 5	?	1	1	1
4 6	?	2	1	2

R1 : IF SFH1=1 and SSP1=1 Then Y(.) = ●

R2 : IF SFH1=2 and SSP1=2 and PT5=1 Then Y(.) = ●

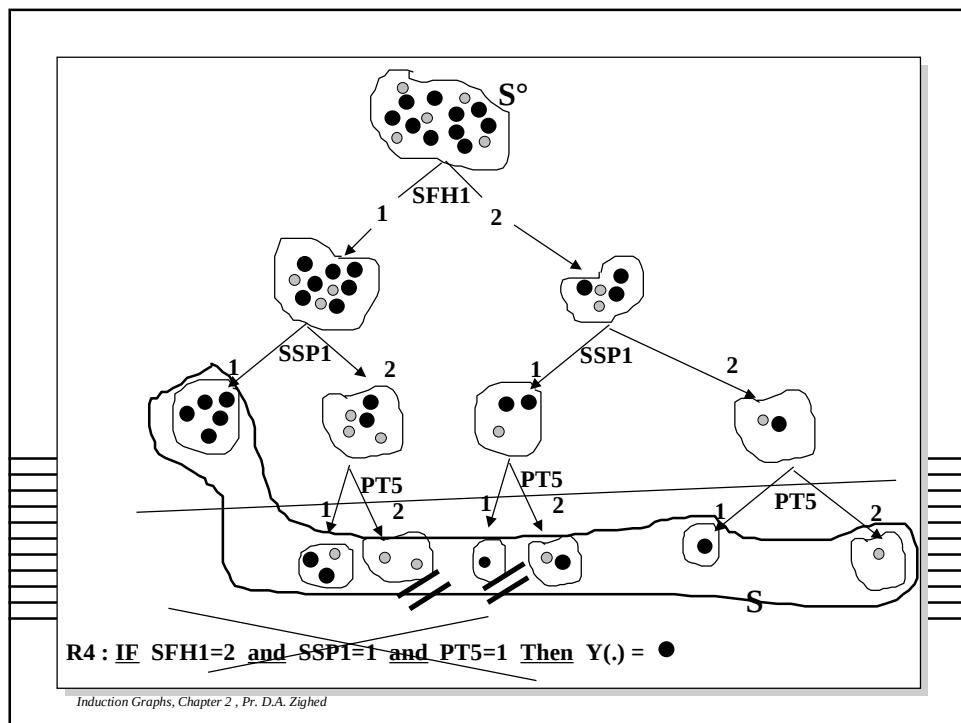
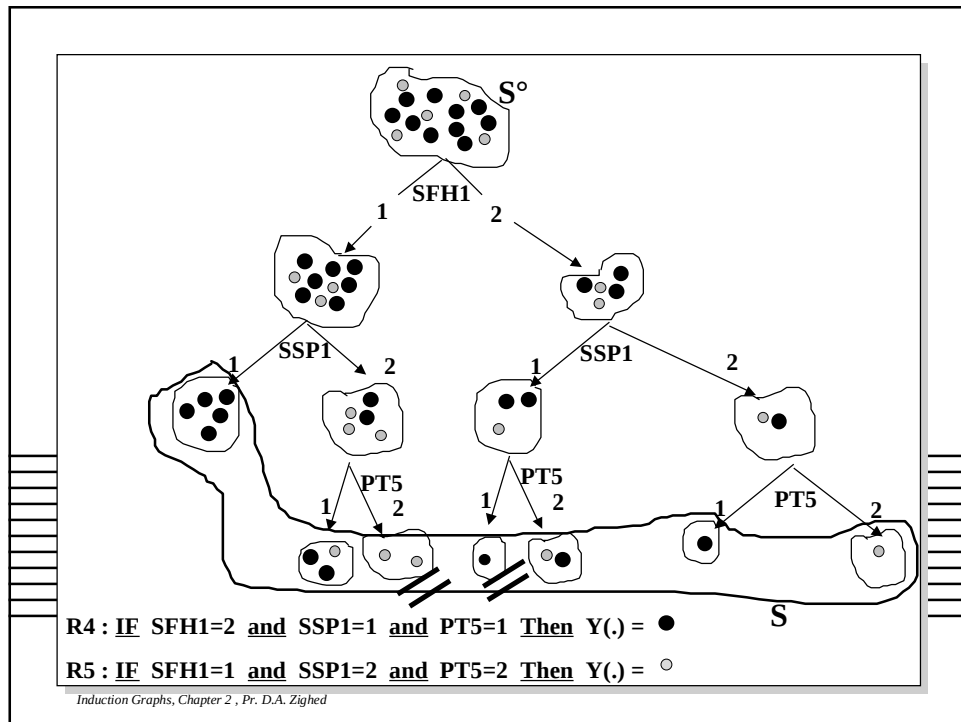
R3 : IF SFH1=2 and SSP1=2 and PT5=2 Then Y(.) = ○

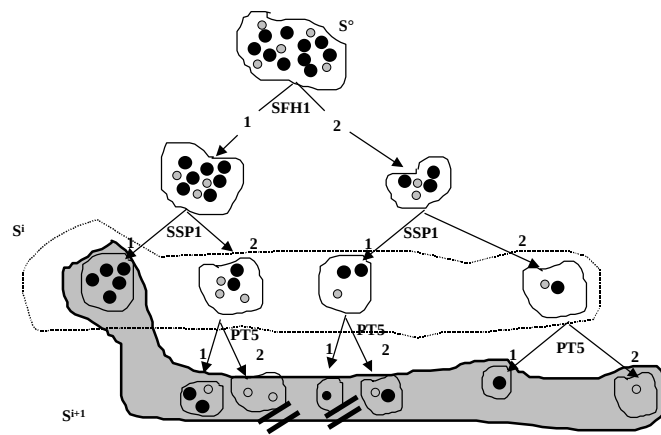
R4 : IF SFH1=2 and SSP1=1 and PT5=1 Then Y(.) = ●

R5 : IF SFH1=1 and SSP1=2 and PT5=2 Then Y(.) = ○

Depth		Predicated cementation
20	By R2	2 Good
21	By R3	1 Bad
32	By R2	2 Good
4 5	By R1	2 Good
4 6	No Rules	? Indetermined

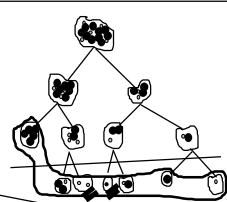
Induction Graphs, Chapter 2 , Pr. D.A. Zighed



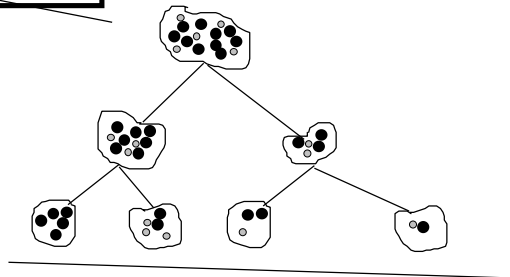


How shall we evaluate the benefit from S^i to S^{i+1} ?

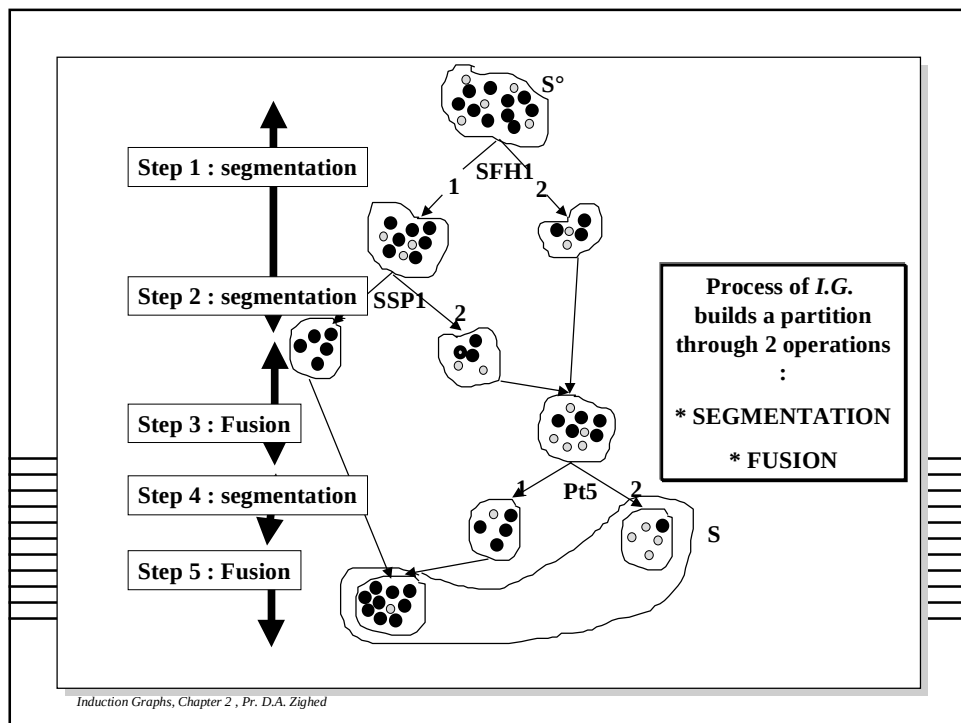
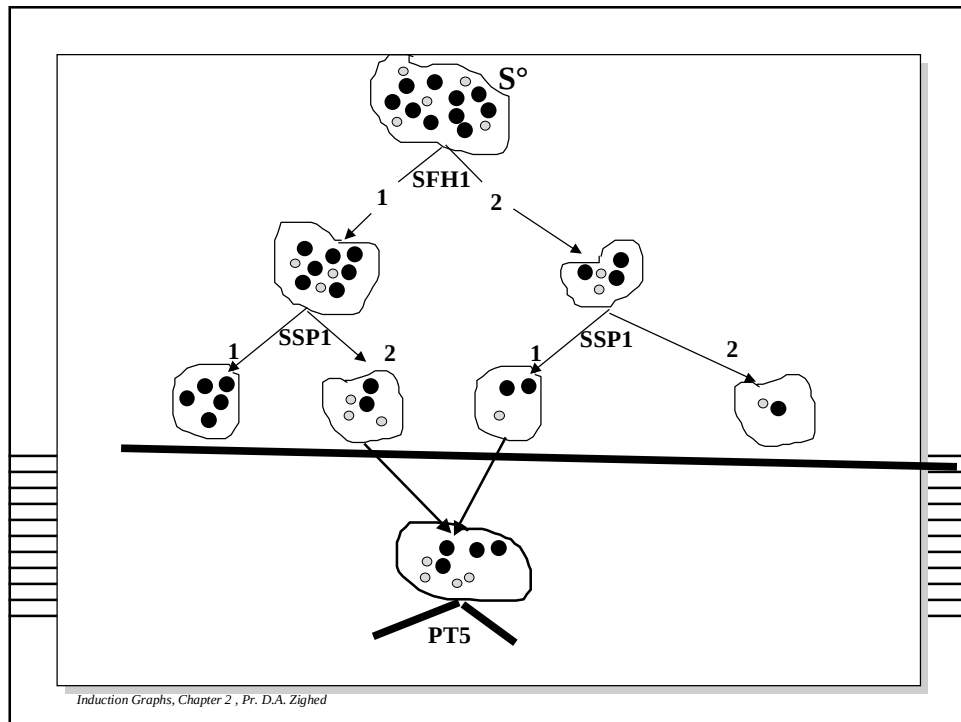
Induction Graphs, Chapter 2 , Pr. D.A. Zighed

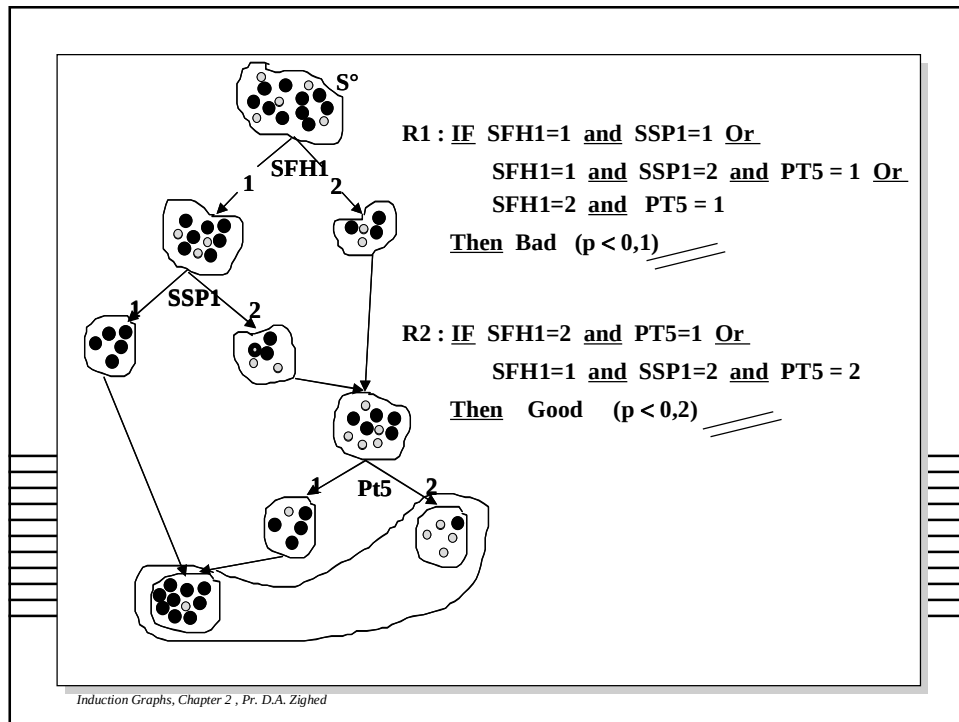


R4 : IE SFH1=2 and SSP1=1 and PT5=1 Then $Y(.) =$

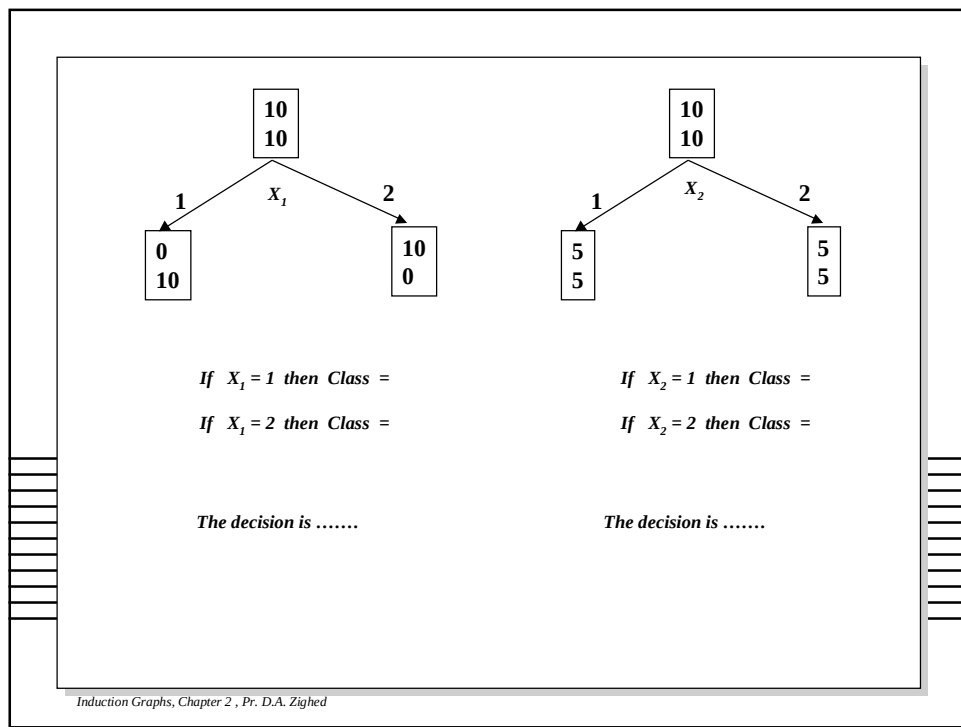
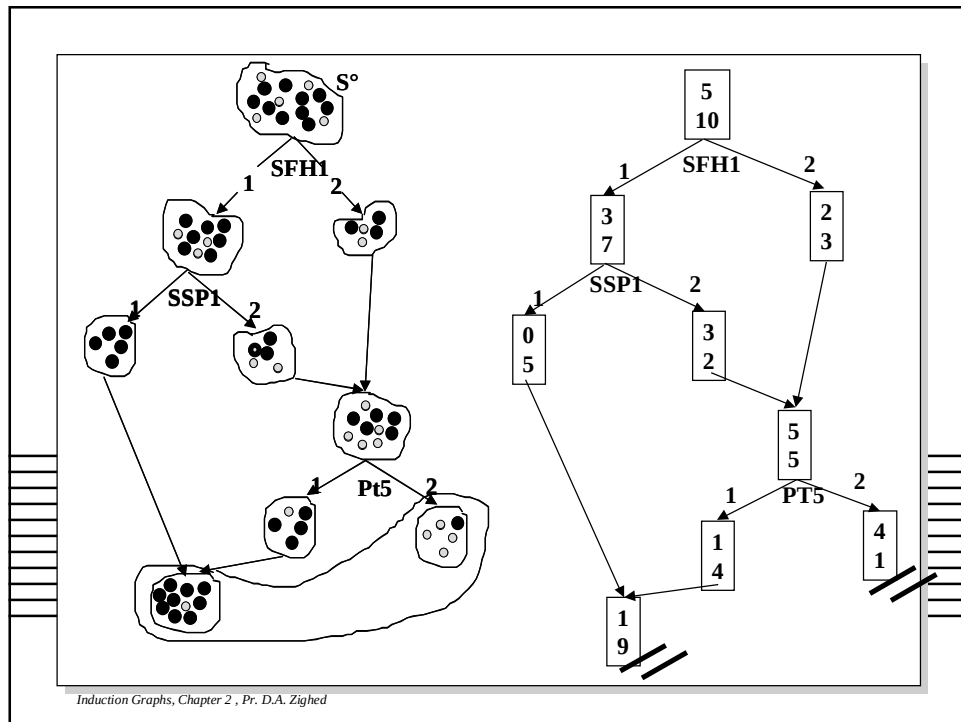


Induction Graphs, Chapter 2 , Pr. D.A. Zighed





- The choice (Segmentation / Fusion)
 - The criteria to move from S^i to S^{i+1}
 - The quantitative variables problem
 - The end of the process
 - The validity of the rules brought about
- Induction Graphs, Chapter 2 , Pr. D.A. Zighed



Entropy measure :

$$h: S \rightarrow \mathbb{R}^+$$

$$S = \bigcup_n S_n$$

$$S_n = \left\{ \gamma \in \mathbb{R}^n / \gamma_i \geq 0 \ (i=1, \dots, n) \text{ et } \sum_{i=1}^n \gamma_i = 1 \right\}$$

➡ **Symetry** $h(\gamma_1, \gamma_2, \dots, \gamma_n) = h(\gamma_{\sigma_1}, \gamma_{\sigma_2}, \dots, \gamma_{\sigma_n})$

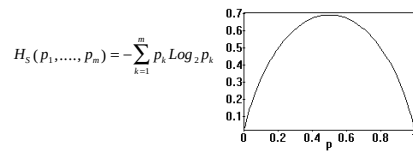
➡ **Minimality** $h(\gamma_1, \gamma_2, \dots, \gamma_n) = 0 \Leftrightarrow \exists i / \gamma_i = 1 \ \forall j \neq i ; \gamma_j = 0$

➡ **Maximality** $h(\gamma_1, \gamma_2, \dots, \gamma_n) = \text{Max} \Leftrightarrow \forall i (i=1, \dots, n) ; \gamma_i = \frac{1}{n}$

➡ **Continuity**

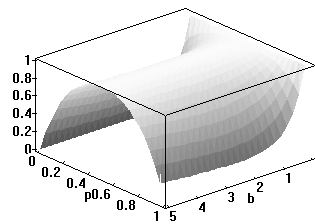
➡ **Concavity**

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

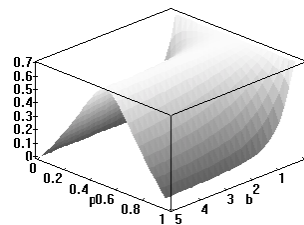


$$H_\beta(p_1, \dots, p_n) = \frac{1}{2^{\beta-1} - 1} \left[\left(\sum_{k=1}^n p_k^\beta \right) - 1 \right] ; \beta > 0; \beta \neq 1$$

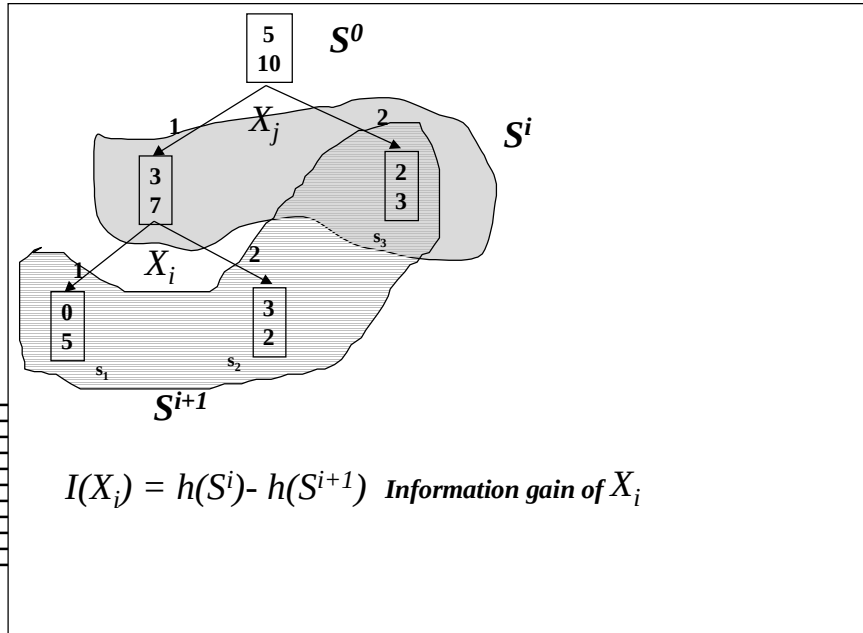
Pour $m=2$ et $n=1$] le graphique se présente ainsi :



$$H^*(p^1, \dots, p^m) = \frac{1}{\alpha} \log \left(\sum_{k=1}^m p_k^\beta \right) ; \beta > 0; \beta \neq 1$$



Induction Graphs, Chapter 2 , Pr. D.A. Zighed



Induction Graphs, Chapter 2, Pr. D.A. Zighed

Segmentation

Initialisations (criteria, stopping condition, parameters...)

Saturation $\leftarrow$ **False**

Repeat

Segmentation $\leftarrow$ **False**

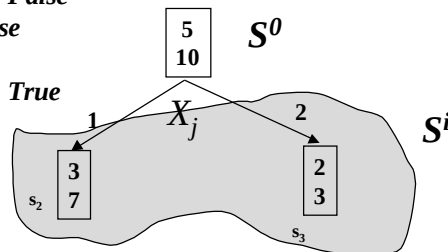
IF Segmentation = **False**

Then

Saturation $\leftarrow$ **True**

End

Until saturation

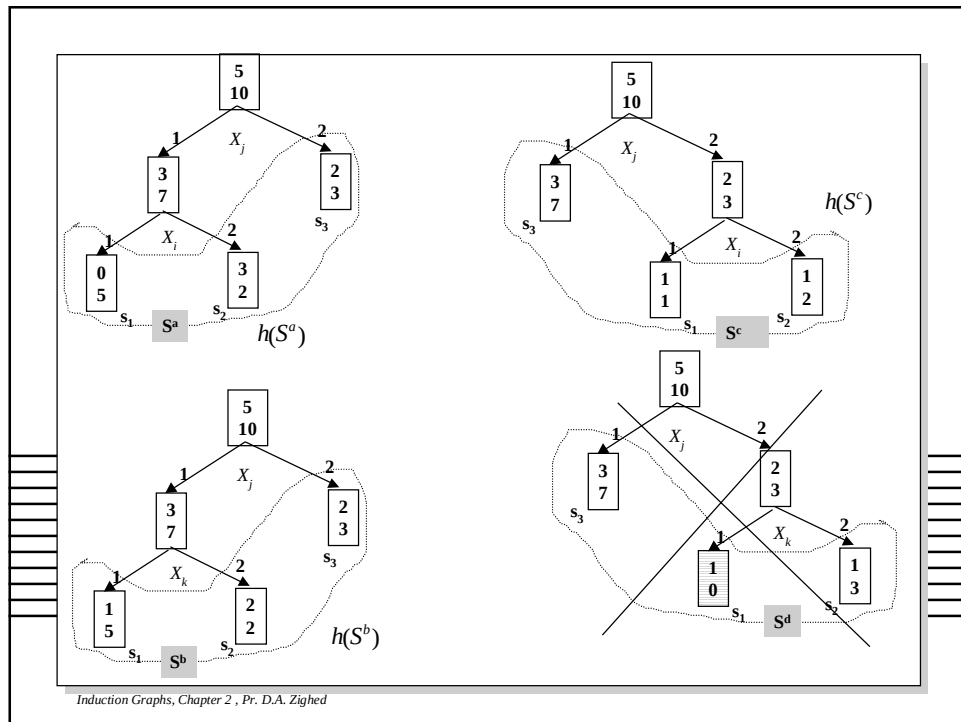


$$h(S^i) = - \sum_{j=1}^2 f(s_j) \sum_{i=1}^2 f(c_i / s_j) \log_2 f(c_i / s_j)$$

Contrainte : $\text{Card}(s_i) > n_0 = 2$

X_i, X_k $\text{Card}(X^{-1}(\Omega)) = 2$

Induction Graphs, Chapter 2, Pr. D.A. Zighed

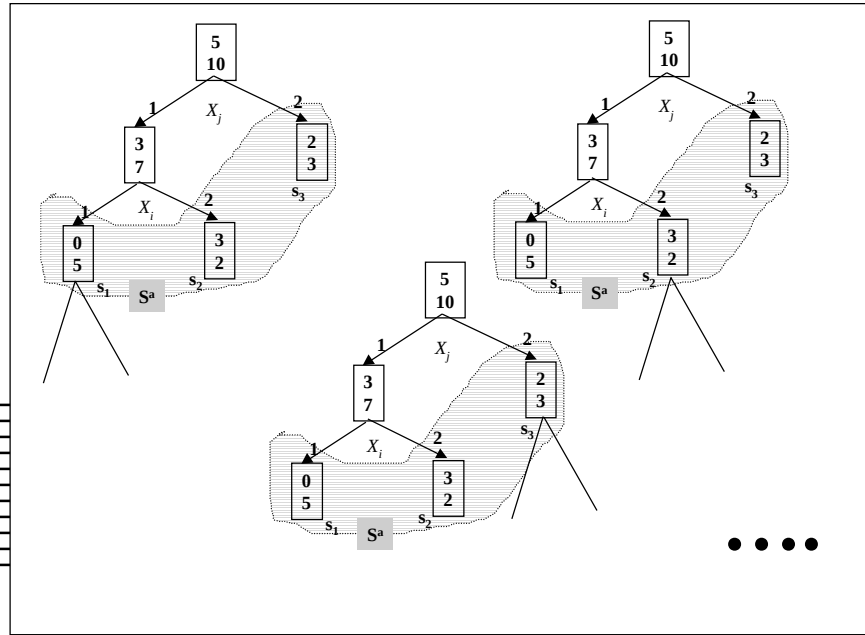


Function Segmentation : Logical

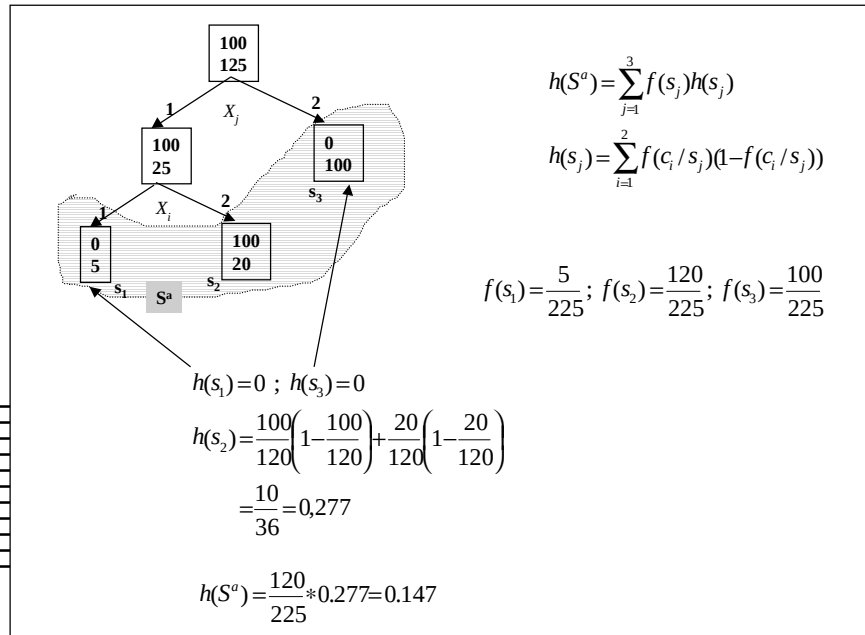
IF $\left(h(S^i) - \min_{j=a,b,c} (h(S^j)) \right) > 0$

Then **Segmentation** $\leftarrow$ **True**
 $\leftarrow$ (node, variable)

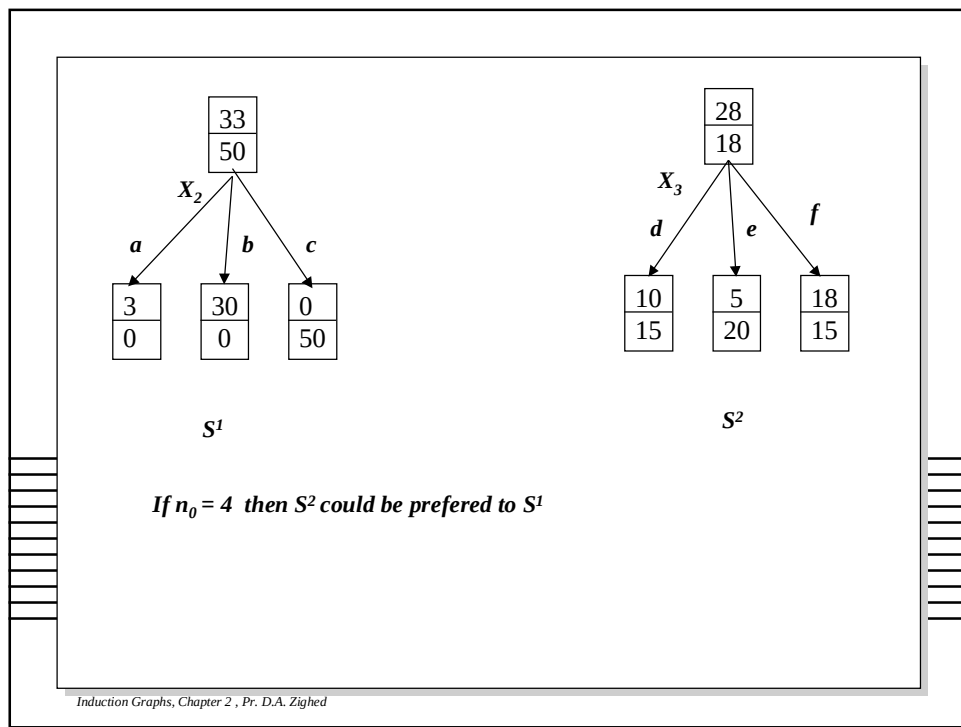
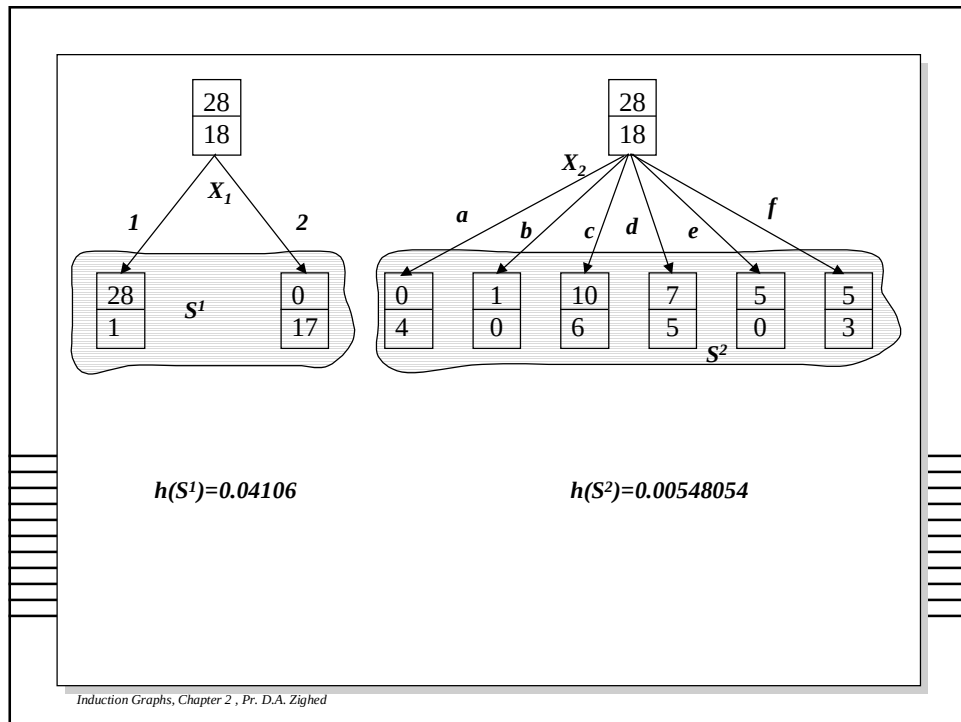
End

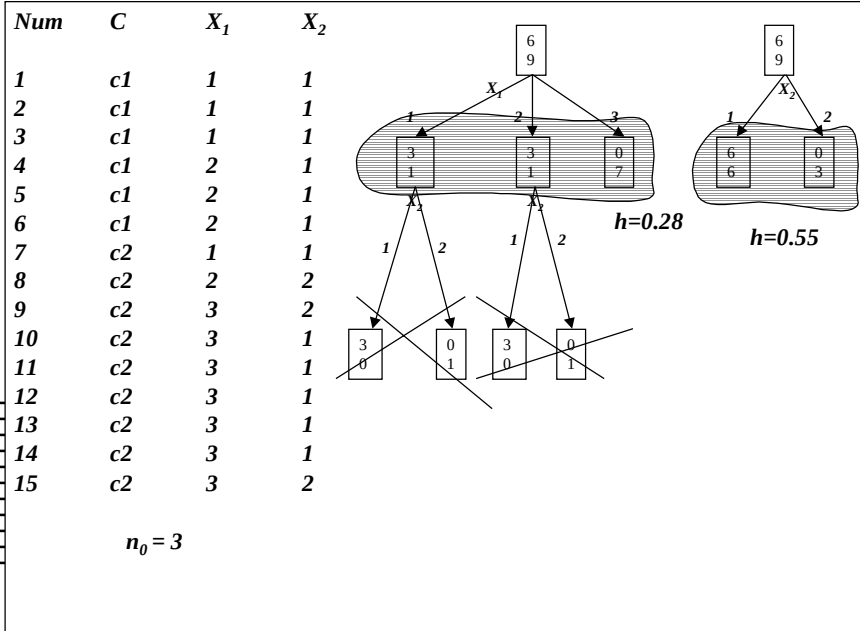


Induction Graphs, Chapter 2 , Pr. D.A. Zighed

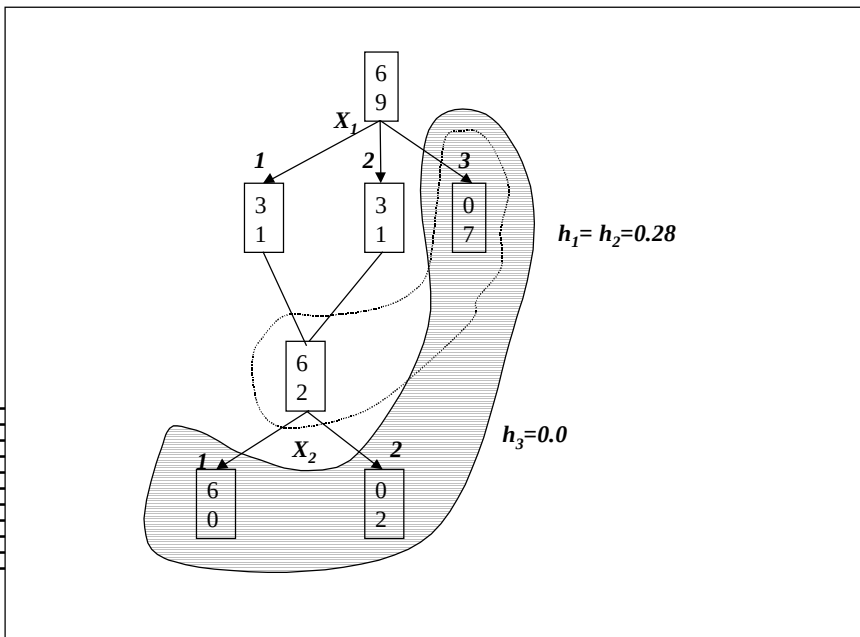


Induction Graphs, Chapter 2 , Pr. D.A. Zighed





Induction Graphs, Chapter 2 , Pr. D.A. Zighed



Induction Graphs, Chapter 2 , Pr. D.A. Zighed

Uncertainty measure on a partition

$$I : P(\Omega, X) \rightarrow R^+$$

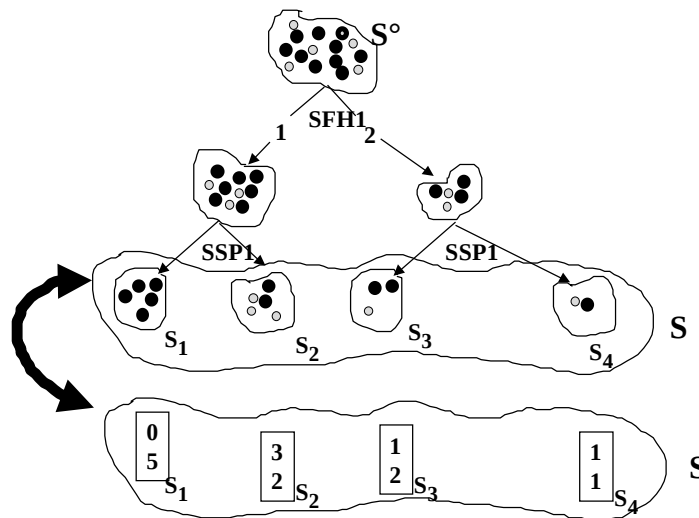
$$S \in P(\Omega, X) \rightarrow I(S)$$

where

$$P(\Omega, X)$$

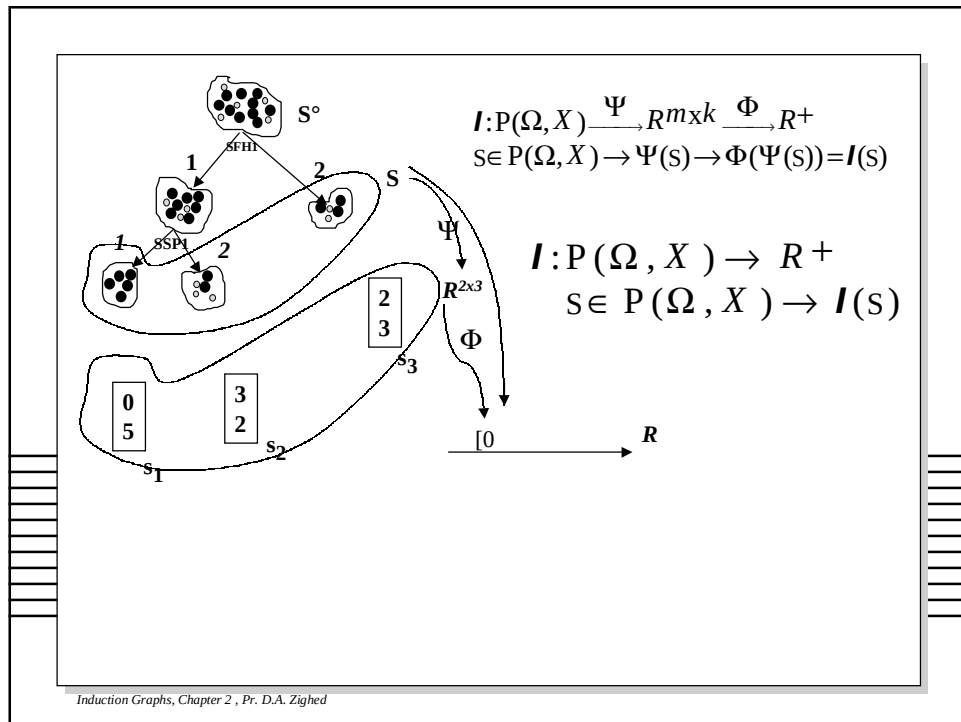
is the set of partitions generated on Ω by the variables $X_1(.), X_2(.), \dots, X_p(.)$

Induction Graphs, Chapter 2, Pr. D.A. Zighed



$S = \{S_1, S_2, S_3, S_4\}$ generated on Ω by $SFH1$ and $SSP1$.

Induction Graphs, Chapter 2, Pr. D.A. Zighed



■ 1_ Symetry

$$S = \left(s_1, s_2, \dots, s_j, \dots, s_k \right) \in R^{m \times k}$$

and σ a permutation

$$S_{\sigma} = \left(s_{\sigma_1}, s_{\sigma_2}, \dots, s_{\sigma_j}, \dots, s_{\sigma_k} \right)$$

Then

$$I(S) = I(S_{\sigma})$$

■ 2 _ Independance

$$S = \left(s_1, s_2, \dots, s_j, \dots, s_k \right) \in R^{m \times k}$$

$$S' = \left(s_1, s_2, \dots, s_{j_0}, s_{j_1}, \dots, s_k \right) \in R^{m \times k}$$

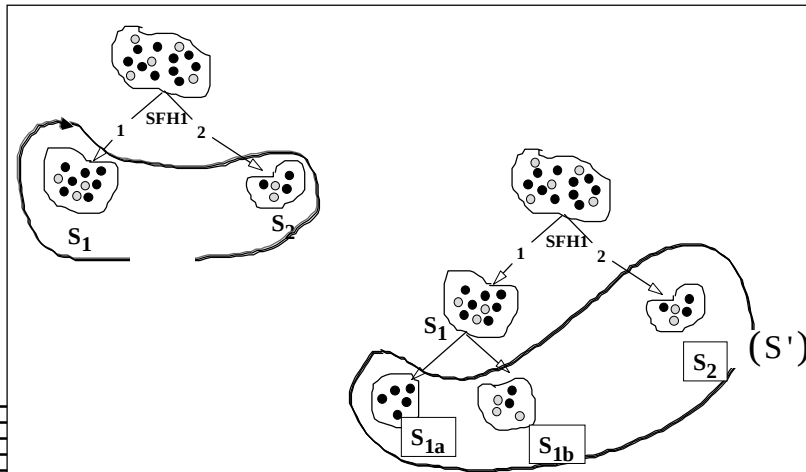
With

$$s_j = s_{j_0} \cup s_{j_1}$$

Then

$$I(S) - I(S') = \varphi(s_{j_0}, s_{j_1})$$

Induction Graphs, Chapter 2 , Pr. D.A. Zighed



$$I(S) - I(S') = \varphi(s_{1a}, s_{1b})$$

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

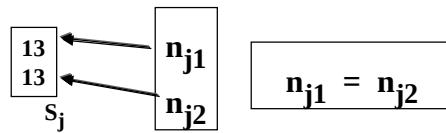
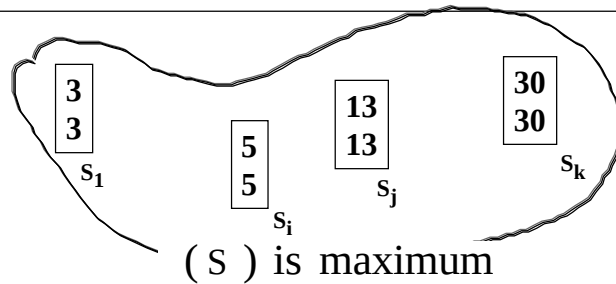
$$s_j = \begin{pmatrix} n_{j0} \\ n_{j1} \\ \vdots \\ n_{jm} \end{pmatrix},$$

If $\forall s_j \in S \quad (j = 1, \dots, k)$

$$n_{j1} = n_{j2} = \dots = n_{jm}$$

$I(S)$ is maximum

Induction Graphs, Chapter 2, Pr. D.A. Zighed



$$n_{11} = n_{12}$$

$$n_{i1} = n_{i2}$$

$$n_{k1} = n_{k2}$$

Induction Graphs, Chapter 2, Pr. D.A. Zighed

■ 4_ Minimality

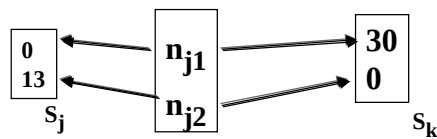
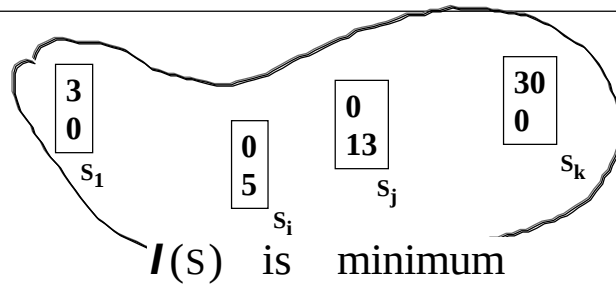
$$s_j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ n_{jp} \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

If $\forall s_j \in S \quad (j=1, \dots, k), \exists p \quad / \quad n_{jp} \neq 0$

and $\forall m \neq p \quad n_{jm} = 0$

Then $I(S)$ is minimum

Induction Graphs, Chapter 2 , Pr. D.A. Zighed



Induction Graphs, Chapter 2 , Pr. D.A. Zighed

■5 _ Sample size sensitivity.

$$S = \left(s_1, s_2, \dots, s_j, \dots, s_k \right) \in R^{m \times k}$$

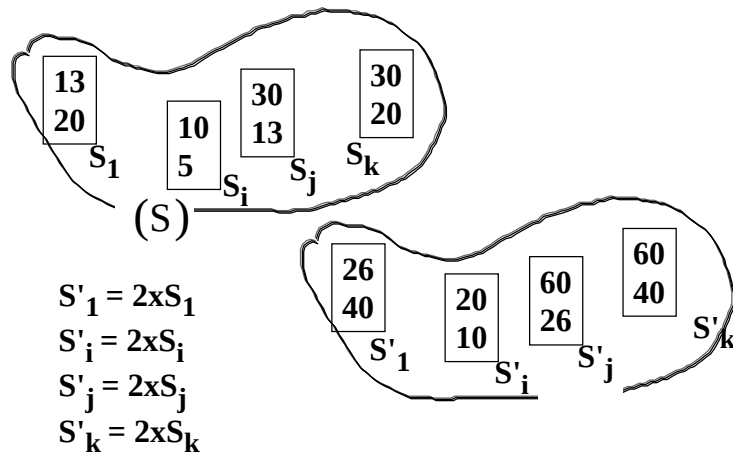
$$S' = \left(n.s_1, n.s_2, \dots, n.s_j, \dots, n.s_k \right) \in R^{m \times k}$$

$$n \geq 1$$

Then

$$(S') \leq I(S)$$

Induction Graphs, Chapter 2 , Pr. D.A. Zighed



Induction Graphs, Chapter 2 , Pr. D.A. Zighed

■ 6 _ Fusion

$$S = (s_1, s_2, \dots, s_{j-1}, s_j, s_{j+1}, \dots, s_{k-1}, s_k) \in R^{m \times k}$$

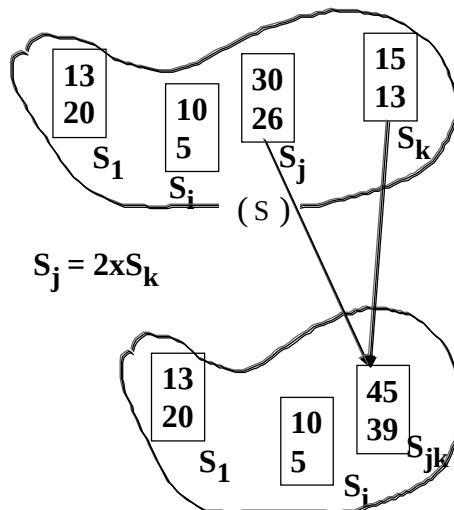
$$\text{If } \exists s_j, s_k \in R^m (j \neq k) / s_j = n \cdot s_k, \quad n \geq 1$$

Then

$$S' = (s_1, s_2, \dots, s_{j-1}, \underbrace{s_j + s_k}_{\text{circled}}, s_{j+1}, \dots, s_{k-1}) \in R^{m \times (k-1)}$$

$$(S') \leq I(S)$$

Induction Graphs, Chapter 2, Pr. D.A. Zighed



Induction Graphs, Chapter 2, Pr. D.A. Zighed

Examples of some uncertainty measures

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

$$S = \left(\begin{array}{ccc} \begin{array}{c} n_{11} \\ n_{21} \\ \vdots \\ n_{m1} \end{array} & \begin{array}{c} n_{21} \\ \vdots \\ \vdots \\ \vdots \end{array} & \cdots & \begin{array}{c} n_{k1} \\ \vdots \\ \vdots \\ n_{mk} \end{array} \end{array} \right) \left(\forall i, j \quad n_{ij} \geq 0 \right)$$

$S_1 \quad - \quad - \quad - \quad S_k$

$$n_{.j} = \sum_{i=1}^m n_{ij}$$

$$n = \sum_{j=1}^k n_{.j} = \sum_{j=1}^k \sum_{i=1}^m n_{ij} \quad (\lambda \geq 0)$$

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

$$I_Q(S) = \sum_{j=1}^k \frac{n_{.j}}{n} \sum_{i=1}^m \frac{n_{ij} + \lambda}{n_{.j} + m\lambda} \left(1 - \frac{n_{ij} + \lambda}{n_{.j} + m\lambda} \right)$$

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

$$I_Q(S) = \sum_{j=1}^k \frac{n_{.j}}{n} \sum_{i=1}^m \frac{n_{ij} + \lambda}{n_{.j} + m\lambda} \left(1 - \frac{n_{ij} + \lambda}{n_{.j} + m\lambda} \right)$$

$$n = 134$$

$$m = 2$$

$$k = 3$$

$$\lambda = 1$$

$$n_{.j} = \sum_{i=1}^m n_{ij}$$

s_1

$$\begin{array}{|c|} \hline 13 \\ 20 \\ \hline 33 \\ \hline \end{array}$$

s_2

$$\begin{array}{|c|} \hline 10 \\ 5 \\ \hline 15 \\ \hline \end{array}$$

s_3

$$\begin{array}{|c|} \hline 60 \\ 26 \\ \hline 86 \\ \hline \end{array}$$

$$\frac{33}{134} \left(\frac{13+1}{33+2*1} \left(1 - \frac{13+1}{33+2*1} \right) + \frac{20+1}{33+2*1} \left(1 - \frac{20+1}{33+2*1} \right) \right) +$$

$$\frac{15}{134} \left(\frac{10+1}{15+2*1} \left(1 - \frac{10+1}{15+2*1} \right) + \frac{5+1}{15+2*1} \left(1 - \frac{5+1}{15+2*1} \right) \right) +$$

$$\frac{86}{134} \left(\frac{60+1}{86+2*1} \left(1 - \frac{60+1}{86+2*1} \right) + \frac{26+1}{86+2*1} \left(1 - \frac{26+1}{86+2*1} \right) \right) =$$

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

$$I_{\mathbf{S}}(S) = \sum_{j=1}^k \frac{n_{\cdot j}}{n} \left[- \sum_{i=1}^m \frac{n_{ij} + \lambda}{n_{\cdot j} + m\lambda} \text{Log} \left(\frac{n_{ij} + \lambda}{n_{\cdot j} + m\lambda} \right) \right]$$

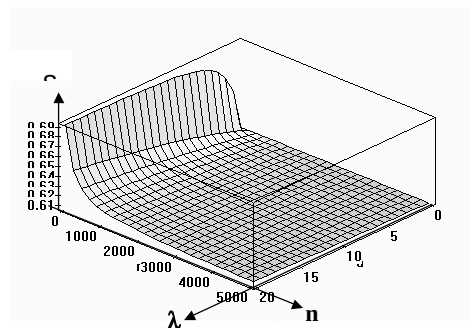
For $\lambda = 0$,

$$I_{\mathbf{S}}(S) = \sum_{j=1}^k \frac{n_{\cdot j}}{n} \left[- \sum_{i=1}^m \frac{n_{ij}}{n_{\cdot j}} \text{Log} \left(\frac{n_{ij}}{n_{\cdot j}} \right) \right]$$

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

$$I_{\mathbf{S}}(S) = \sum_{j=1}^k \frac{n_{\cdot j}}{n} \left[- \sum_{i=1}^m \frac{n_{ij} + \lambda}{n_{\cdot j} + m\lambda} \text{Log} \left(\frac{n_{ij} + \lambda}{n_{\cdot j} + m\lambda} \right) \right]$$

S_1	S_2
0.1	0.4
0.3	0.2



Induction Graphs, Chapter 2 , Pr. D.A. Zighed

$$I_D(s) = \frac{1}{2^{\beta-1}-1} \sum_{j=1}^k \frac{n_{\cdot j}}{n} \left(\sum_{i=1}^m \left(\frac{n_{ij} + \lambda}{n_{\cdot j} + m\lambda} \right)^{\beta} - 1 \right)$$

$$(\beta \in [0 ; 1])$$

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

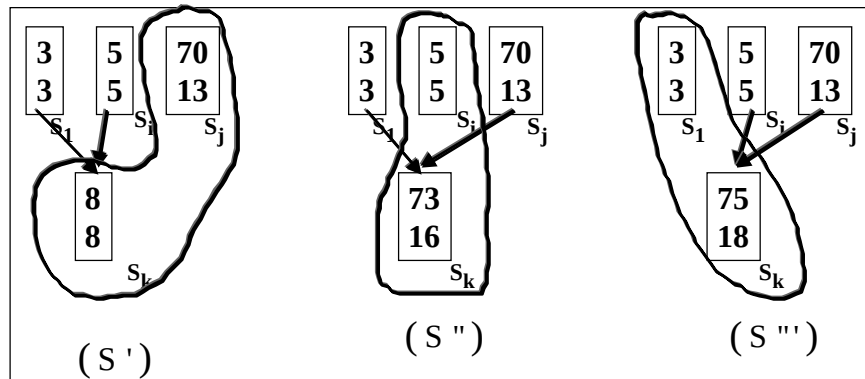
Algorithm for I.G.

```

Initialisation
Saturation <---- false
Repeat
  Fusion <---- false
  Segmentation <---- false
  IF Fusion Or Segmentation
  Then
    IF Merge >= Split
    Then
      Make_fusion
    Else
      Make_split
    End
  Else
    Saturation <---- true
  End
Until saturation

```

Induction Graphs, Chapter 2 , Pr. D.A. Zighed



Function Fusion : Logical

IF $(I(S) - \text{Min}(I(S'), I(S''), I(S''')) > 0$

Then

Fusion $\leftarrow \text{true}$

Merge $\leftarrow (I(S) - \text{Min}(I(S'), I(S''), I(S'''))$

End

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

Function Segmentation : Logical

IF $\left(I(S) - \text{Min}_{j=1, \dots, 3p} (I(S^j)) \right) > 0$

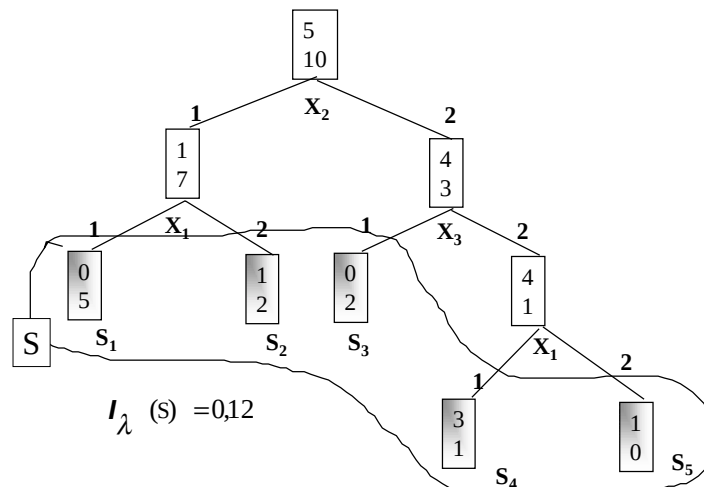
Then

Segmentati **on** $\leftarrow \text{true}$

Split $\leftarrow \left(I(S) - \text{Min}_{j=1, \dots, 3p} (I(S^j)) \right)$

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

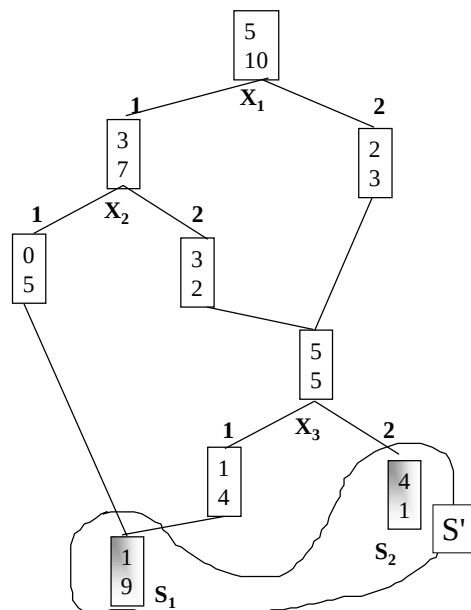
$$I_S(S) = \sum_{j=1}^k \frac{n_{.j}}{n} \left(- \sum_{i=1}^m \frac{n_{ij} + \lambda}{n_{.j} + m\lambda} \log \left(\frac{n_{ij} + \lambda}{n_{.j} + m\lambda} \right) \right) ; \lambda = 0$$



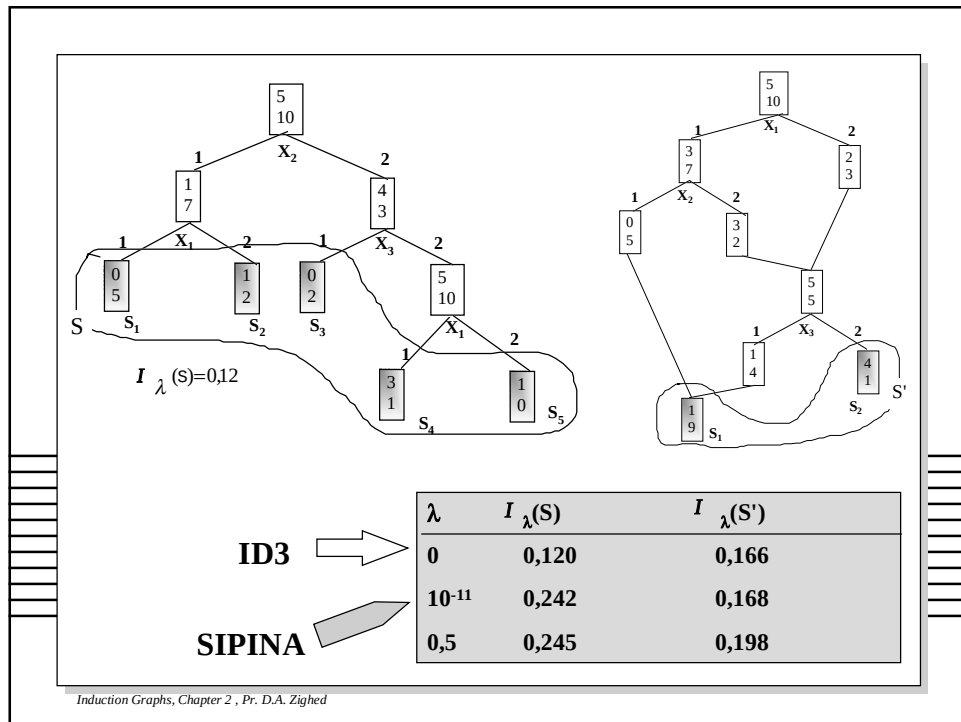
Induction Graphs, Chapter 2 , Pr. D.A. Zighed

$$\lambda = 10^{-10}$$

$$I_{\lambda}(S') = 0,198$$



Induction Graphs, Chapter 2 , Pr. D.A. Zighed



Fixation du paramètre λ

τ Une contrainte d'admissibilité

u

τ
0
:
:
0

$n_u = \tau$

v

$\tau+1$
0
:
:
0

$n_v = \tau + 1$

$$I(u) - I(v) = \max \{ I(u) - I(v) \}$$

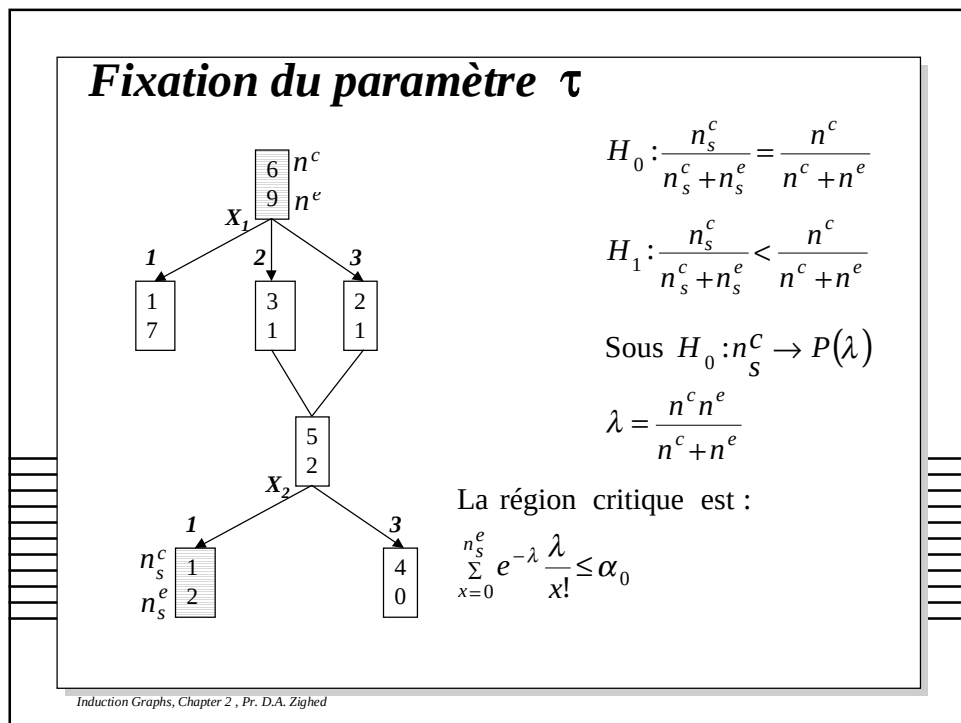
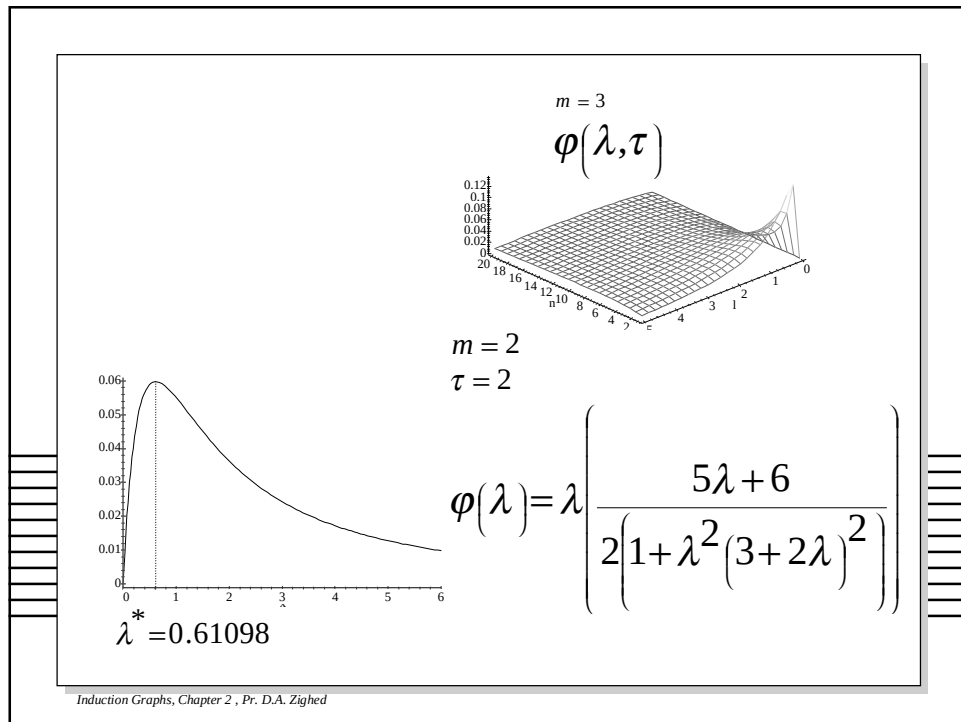
$$\phi(\lambda) = I_{\lambda}(u) - I_{\lambda}(v)$$

$$= \frac{(m-1)^{2\tau^2+2\tau+m\lambda+2\tau m\lambda}}{(\tau+1)^2(\tau+1)^2}$$

Trouver $\lambda^* / \phi(\lambda)$ soit maximum

Trouver $\lambda / \partial \phi(\lambda) = 0$

Induction Graphs, Chapter 2, Pr. D.A. Zighed



Pour un seuil critique $\alpha_0=0.05$, on détermine le nombre d'individus total en n_s tel que si $n_{sc}=0$ on conclut au rejet de H_0 .

D'où on établit :

$$e^{-\lambda} = \alpha_0$$

$$\tau = n_s = -\ln(\alpha_0) \frac{n^c + n^e}{n^c}$$

Wisconsin Breast Cancer Database (January 8, 1991)

Sources:

- Dr. William H. Wolberg (physician)
University of Wisconsin Hospitals
Madison, Wisconsin
USA
- Donor: Olvi Mangasarian (mangasarian@cs.wisc.edu)
Received by David W. Aha (aha@cs.jhu.edu)
- Date: 15 July 1992

Les données

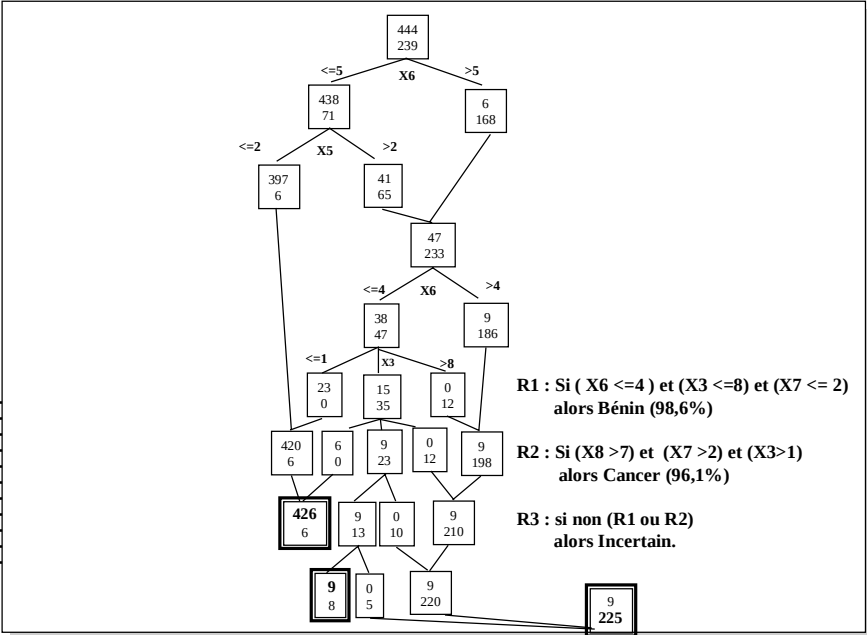
#	Attribute	Domain
1.	Sample code number	id number
2.	Clump Thickness	1 - 10
3.	Uniformity of Cell Size	1 - 10
4.	Uniformity of Cell Shape	1 - 10
5.	Marginal Adhesion	1 - 10
6.	Single Epithelial Cell Size	1 - 10
7.	Bare Nuclei	1 - 10
8.	Bland Chromatin	1 - 10
9.	Normal Nucleoli	1 - 10
10.	Mitoses	1 - 10
11.	Class:	(2 for benign, 4 for malignant)

Echantillon : 683

Bénin : 444

Cancer : 239

Induction Graphs, Chapter 2 , Pr. D.A. Zighed



Induction Graphs, Chapter 2 , Pr. D.A. Zighed

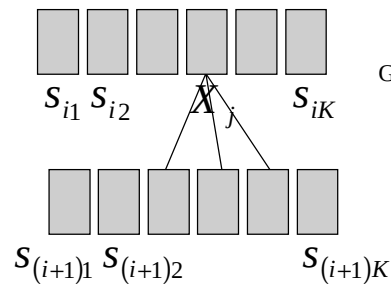
Others Induction Graph methods

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

C4.5 Quinlan 1993

$$S_i = \{s_{i1}, s_{i2}, \dots, s_{ik}\} \text{ and } h(S_i) = \sum_{j=1}^K f(s_{ij}) \sum_{t=1}^m -f(c_t / s_{ij}) \log f(c_t / s_{ij})$$

$$S_{i+1} = \{s_{(i+1)1}, s_{(i+1)2}, \dots, s_{(i+1)L}\} \text{ and } h(S_{i+1})$$



$$\text{Gain Ratio: } I(X_j) = \frac{h(S_i) - h(S_j)}{-\sum_{j=1}^L f(s_{(i+1)j}) \ln f(s_{(i+1)j})}$$

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

Strategy : segmentation

admissibility constraint : 2

Stopping rule : Chi-2 test

Discretization : binary

prunning rule : lower limit of the confidence interval

Induction Graphs, Chapter 2 , Pr. D.A. Zighed

CART : Classification And Regression Trees
Breiman et al. 19984

Strategy : Gini or Twoing

Strategy : binary segmentation

admissibility constraint : 2

Stopping rule : None

Discretization : binary

prunning rule : Complexity cost criteria

Induction Graphs, Chapter 2 , Pr. D.A. Zighed