RAONAMENT BASAT EN L'EXPERIÈNCIA
Raonament Basat en Casos

(SBC-CBR Part IV – Problemes en el Desenvolupament de Sistemes CBR)

Miquel Sànchez-Marrè

Intelligent Data Science and Artificial Intelligence Research Centre (IDEAI-UPC)

Knowledge Engineering and Machine Learning Group (KEMLG-UPC)

Computer Science Dept.
Universitat Politècnica de Catalunya · Barcelona**Tech**

miquel@cs.upc.edu
http://www.cs.upc.edu/~miquel

Course 2023/2024

https://kemlg.upc.edu

PART 4 – PROBLEMES EN EL DESENVOLUPAMENT DE SISTEMES DE CBR

PROBLEMES EN EL DESENVOLUPAMENT DE SISTEMES DE CBR

PROBLEMES EN SISTEMES CBR JERÀRQUICS

- **Competència**
  - Valoració de la similitud
  - Informació desconeguda (valors que falten)
  - Discretització
  - Cerca infructuosa / Pèrdua de casos en estructures jeràrquiques
- **Eficiència en Temps**
  - Degradació de les estructures jeràrquiques
  - Overhead d'aprenentatge
- **Eficiència en Espai**
  - Quan és necessari aprendre un cas nou?
  - Hem d'oblidar alguns casos?

Competència
Valoració de la similitud

Valoració de la similitud

- Algoritme Nearest Neighbour (NN)

$$\text{Full-dist}\,(C_i,C_j) = \sum_{k=1}^{n} w_k * \text{atr-dist}\,(C_{ik},C_{jk}) \;/\; \sum_{k=1}^{n} w_k$$

- **Problemes**:
  - **Es perd la rellevància dels atributs** quan n augmenta
  - **La majoria només utilitzen valors quantitatius**

Distància l'Eixample: heterogènia sensible al pes

[Sànchez-Marrè, 1996; Sànchez-Marrè et al., 1998]

$$d(C_i, C_j) = \frac{\sum_{k=1}^{n} e^{w_k} \times d(A_{ki}, A_{kj})}{\sum_{k=1}^{n} e^{w_k}}$$

$$d(A_{ki}, A_{kj}) = \begin{cases} \dfrac{|quantval(A_{ki}) - quantval(A_{kj})|}{upperval(A_k) - lowerval(A_k)} & \text{if } A_k \text{ is an ordered attribute and } w_k \le \alpha \\ \dfrac{|qualval(A_{ki}) - qualval(A_{kj})|}{\#\text{mod}(A_k) - 1} & \text{if } A_k \text{ is an ordered attribute and } w_k > \alpha \\ 1 - \delta_{qualval(A_{ki}), qualval(A_{kj})} & \text{if } A_k \text{ is a non-ordered attribute} \end{cases}$$

Proves experimentals (1)

- $$d(C_i,C_j) = \left(\sum_{k=1}^{n} weight^r * |d(A_{ki},A_{kj})|^r \Big/ \sum_{k=1}^{n} weight^r \right)^{1/r}$$

Similarity Measure	r	Weight	Ordered attributes $d(A_{ki}, A_{kj})$	
			$W_k > \alpha$	$W_k \leq \alpha$
Discrete Manhattan (MD)	1	W_k	$\lvert qlv(A_{ki}) - qlv(A_{kj})\rvert / (\#mod(A_k) - 1)$ **[a]**	
Discrete Euclidean (ED)	2	W_k	**[a]**	
Discrete Exp.-weighted Manhattan (EMD)	1	e^{Wk}	**[a]**	
Continuous Manhattan (MC)	1	W_k	$\lvert qtv(A_{ki}) - qtv(A_{kj})\rvert / \lvert(upv(A_k) - lowv(A_k)\rvert$ **[b]**	
Continuous Euclidean (EC)	2	W_k	**[b]**	
Continuous Exp.-weighted Manhattan (EMC)	1	e^{Wk}	**[b]**	
Weight-sensitive Manhattan (MW)	1	W_k	**[a]**	**[b]**
Weight-sensitive Euclidean (EW)	2	W_k	**[a]**	**[b]**
Weight-sensitive Exp.-weighted Manhattan (EIX)	1	e^{Wk}	**[a]**	**[b]**

Proves experimentals (2)

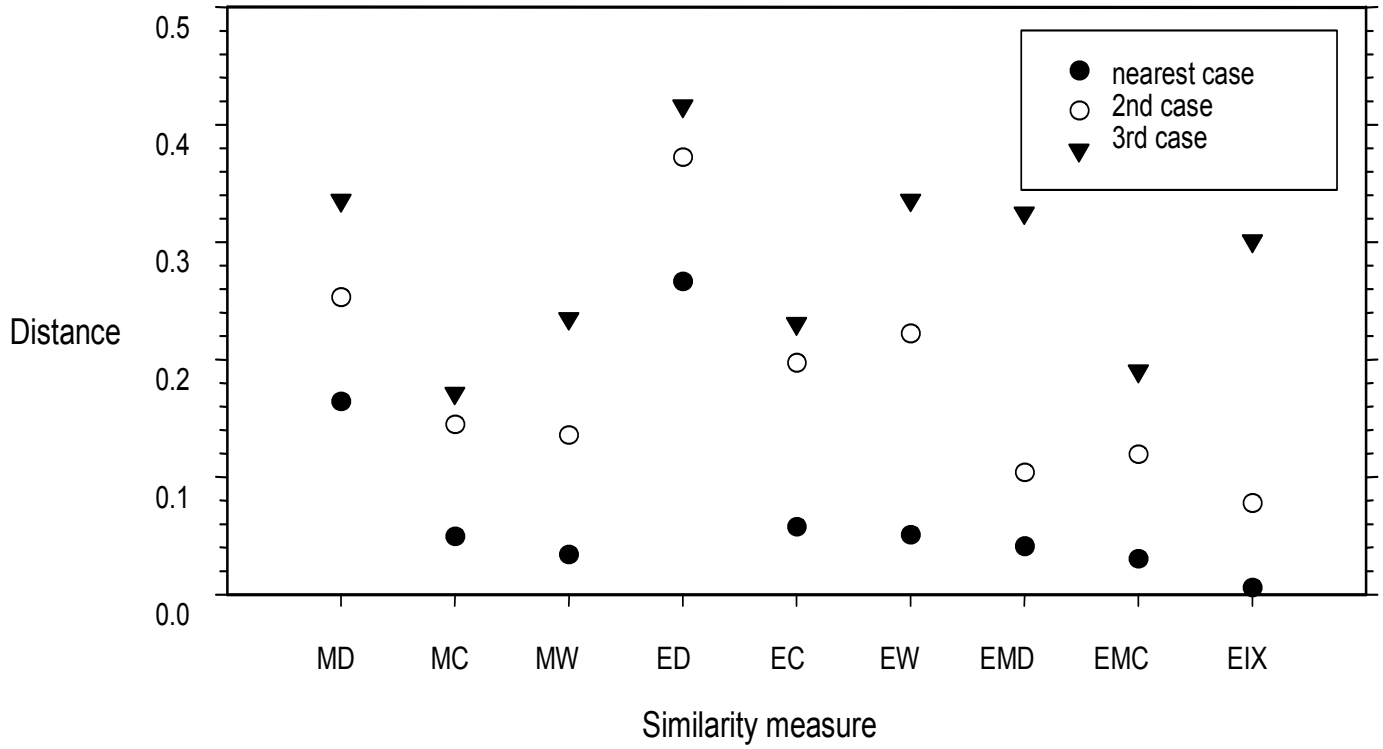
- **EDAR de Girona**
  - 45.000 m3/dia - 70.000 hab.
  - 396 casos reals del període 1995/1996
- **EDAR de Lloret**
  - 13.000 m3/dia - 20.000 hab. (hivern)
  - 45.000 m3/dia - 150.000 hab. (estiu)
  - 234 casos reals del període 1996/1997
- **Conjunt d'entrenament de 10 *batches***
- **10 mesures de semblança provades** (100 taules de recuperació)
- **Biblioteques de casos inicialitzades amb casos representatius**

Recuperació òptima

Similarity Measure	% Optimal retrieval Lloret's WWTP	% Optimal retrieval Girona's WWTP
Discrete Manhattan (MD)	60	60
Discrete Euclidean (ED)	80	60
Discrete Exponential-weighted Manhattan (EMD)	80	70
Continuous Manhattan (MC)	50	90
Continuous Euclidean (EC)	60	80
Continuous Exponential-weighted Manhattan (EMC)	60	70
Weight-sensitive Manhattan (MW)	60	90
Weight-sensitive Euclidean (EW)	80	90
Weight-sensitive Exponential-weighted Manhattan (EIX)	80	90

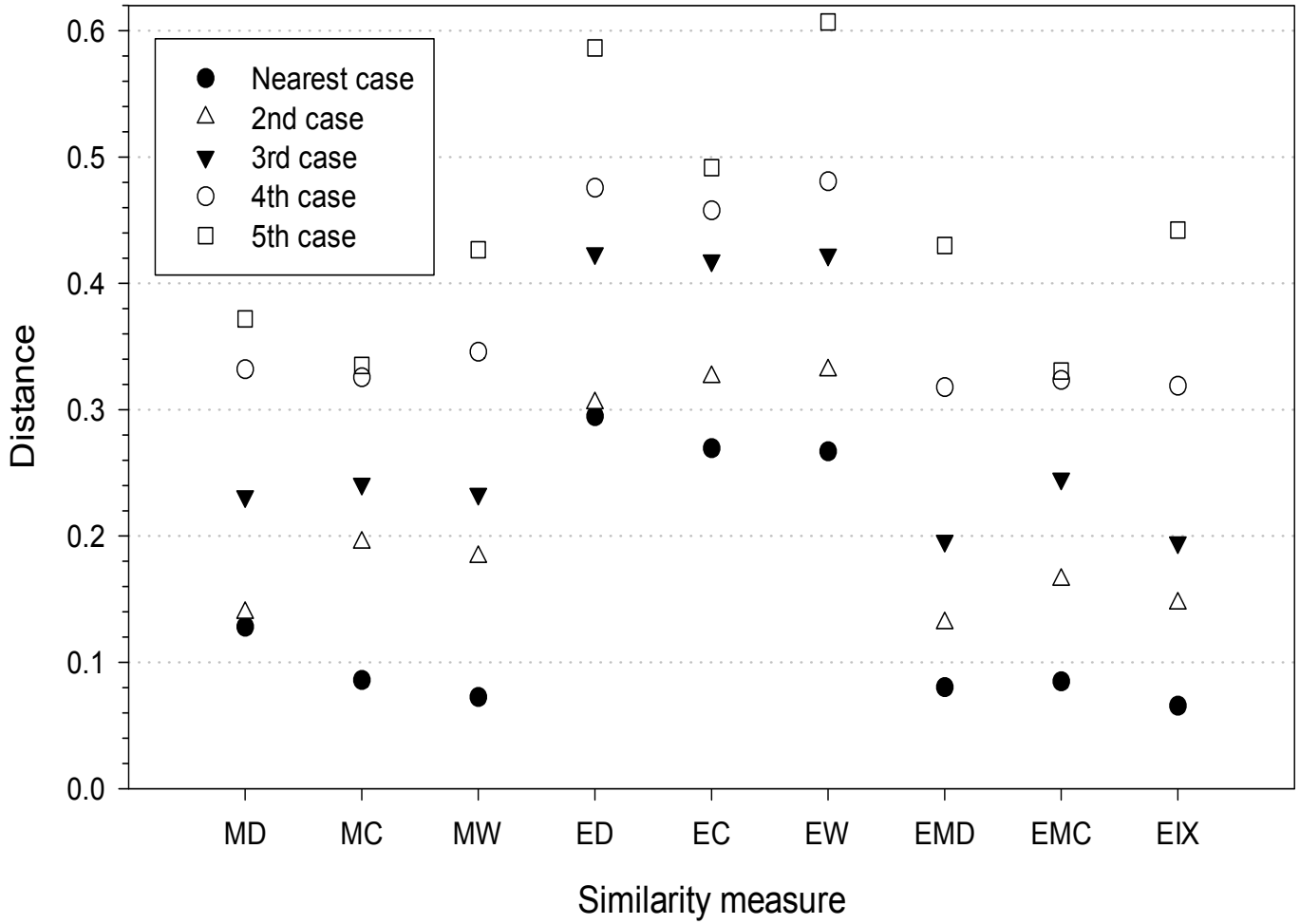
Rànquing de recuperació òptima (1)

- **EDAR de Girona**



Rànquing de recuperació òptima (2)

- **EDAR de Lloret**



Conclusions de l'avaluació de la similitud

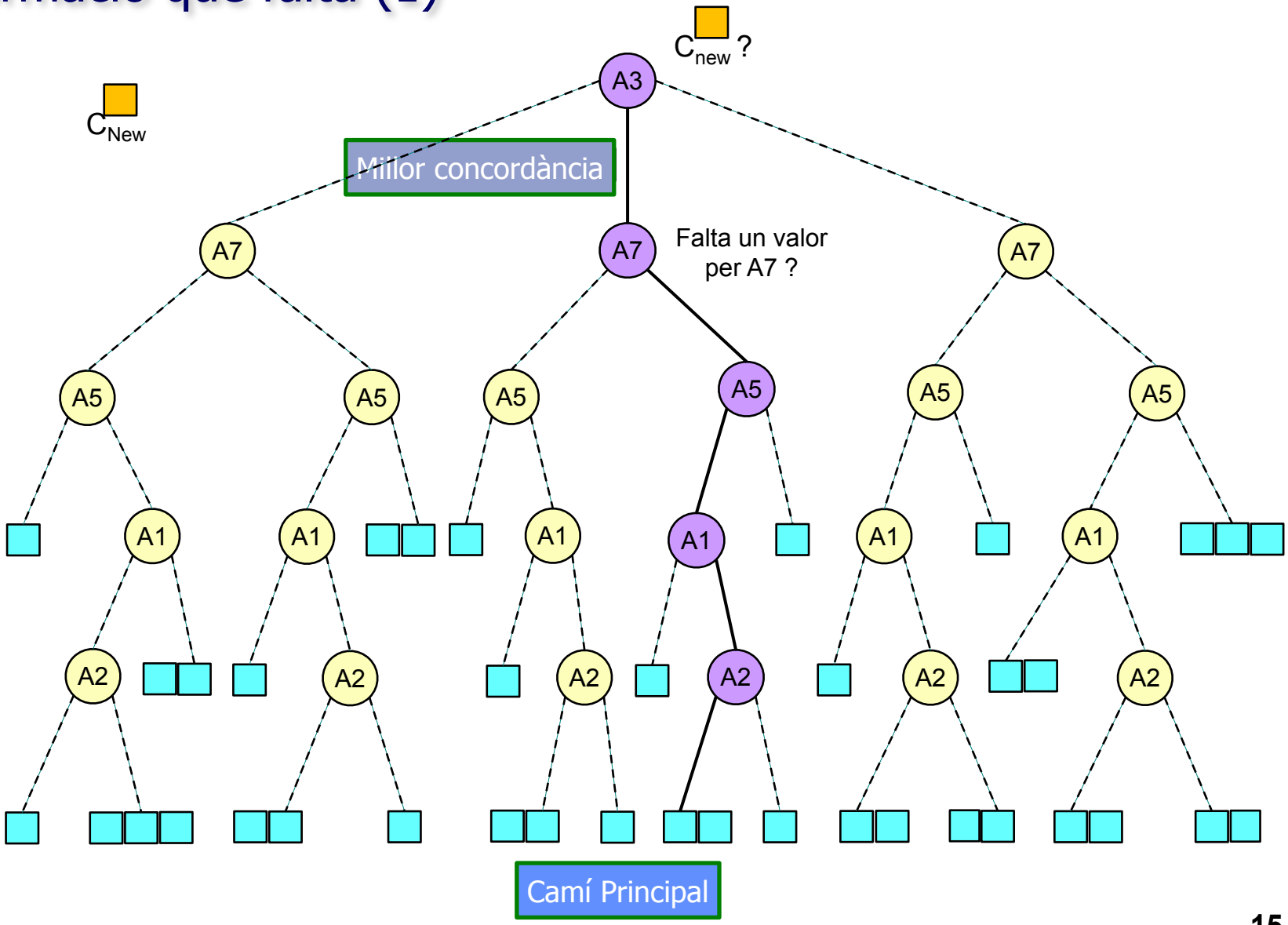
- **Les distàncies contínues i euclidianes** resulten més adequades per a dominis amb **molts valors quantitatius ordenats** i pocs valors categòrics no ordenats
- **Les distàncies Manhattan** resulten més adequades per a dominis amb **categòrics ordenats**
- La distància l'**Eixample** es deriva de la Manhattan i **combina atributs continuous i discrets**. Sembla millór, però és molt **sensible** als **pesos** i al procés de **discretització**
- La selecció de funcions i la ponderació de les funcions són reptes importants

Competència

Informació que falta

[Sànchez-Marrè et al., 1997]

Informació que falta (1)



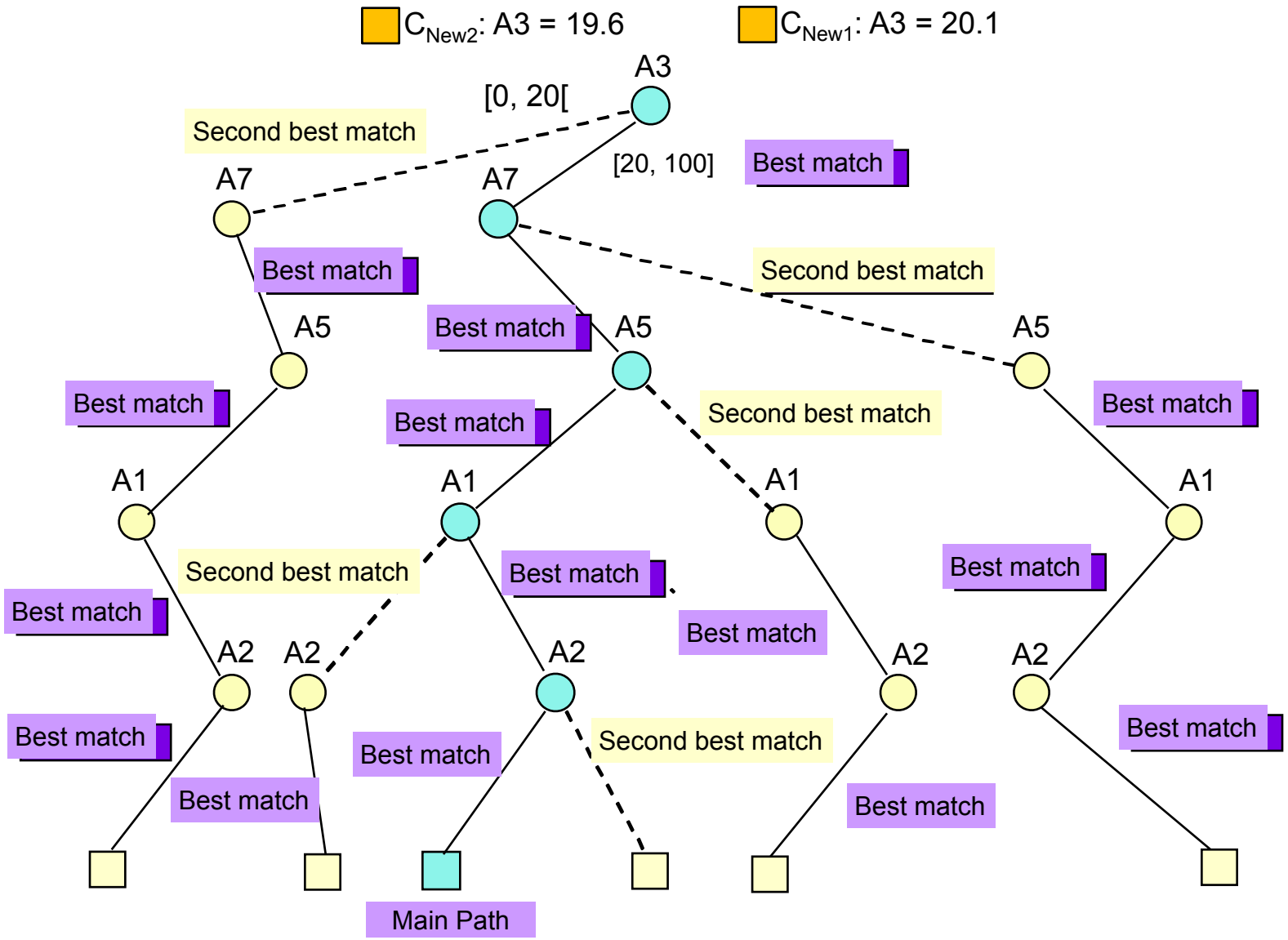
Informació que falta (1)

- **Possibles solucións**:
  - **Aturat la cerca**
  - **Cercar a totes les branques de la jerarquia**
  - **Cercar en altres jerarquies amb diferents ordres de característiques (arbres de discriminació redundants)**
  - **Cercar a la branca més prometedora**
- **Aproximació escollida**:
  - **Associar valors de freqüència a cada branca**
  - **Cercar a la branca amb la freqüència més alta**

Competència
Problemes amb la discretització

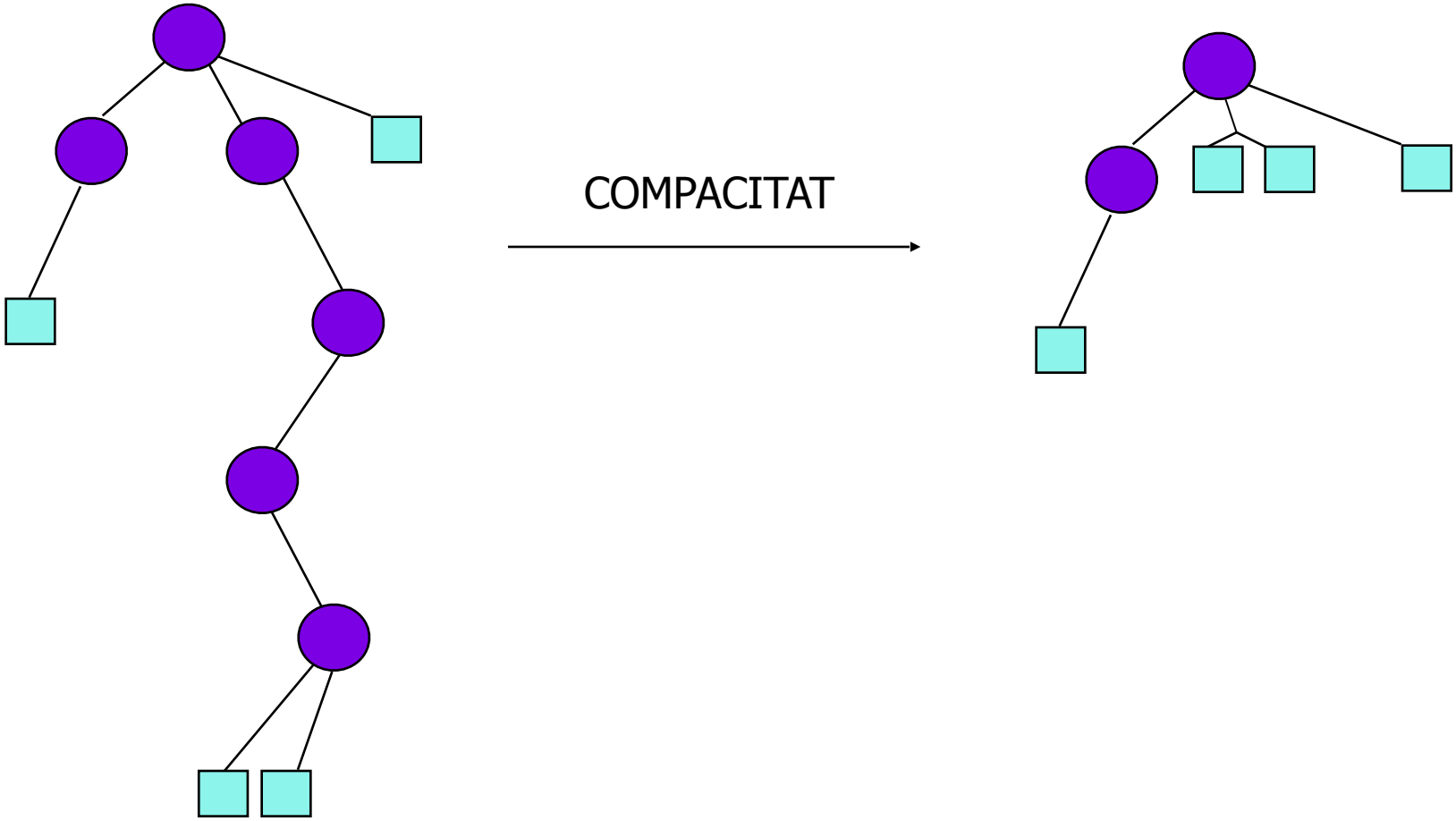
[Sànchez-Marrè et al., 1997]

Discretització: concordança parcial



Temps
Degeneració de la jerarquia

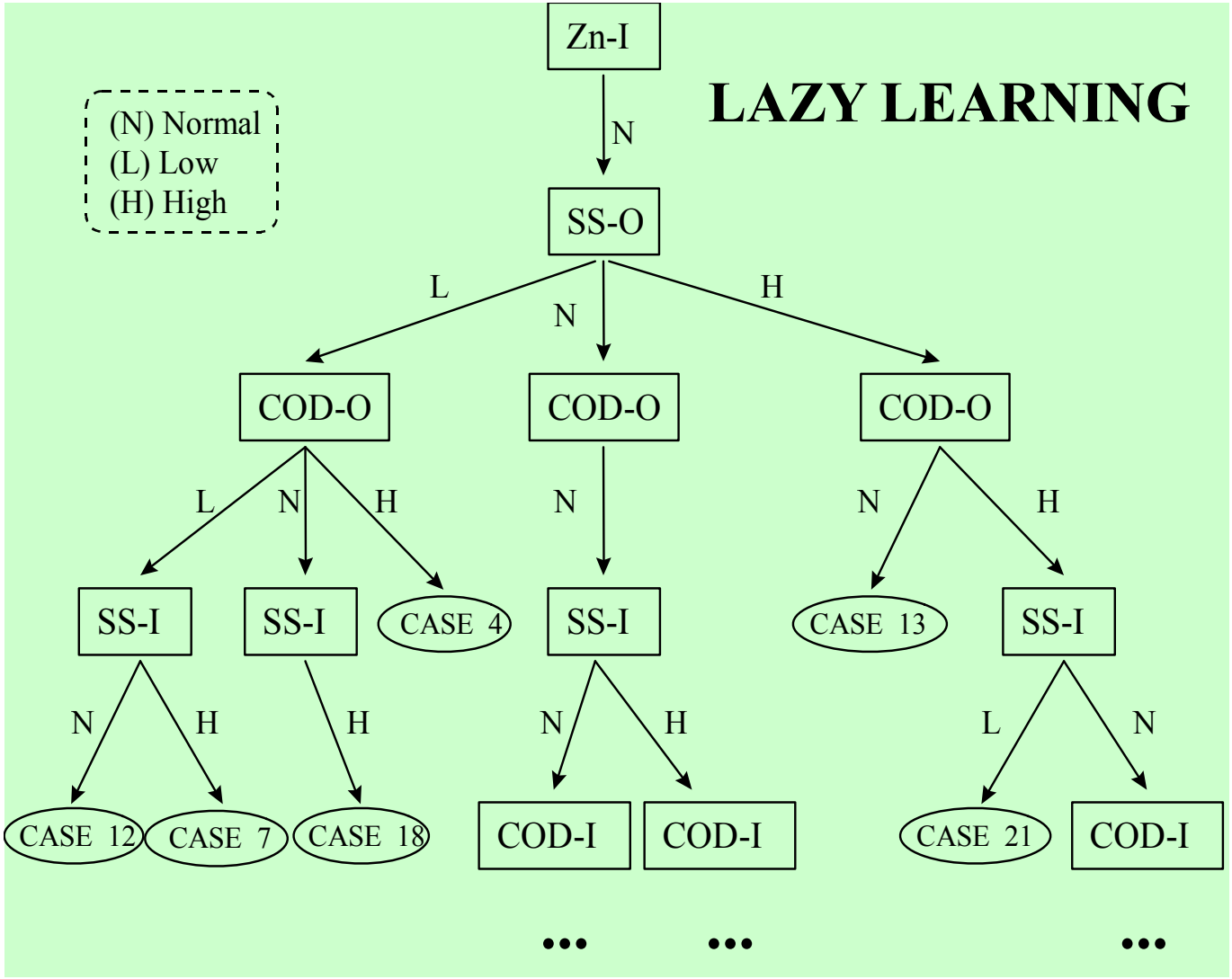
Temps:Degeneració de la jerarquia



Temps

Sobrecàrrega ("Overhead") en l'aprenentatge

Temps: Sobrecàrrega ("Overhead") en l'aprenentatge



Espai
Quan aprendre?
[Sànchez-Marrè et al., 1997]

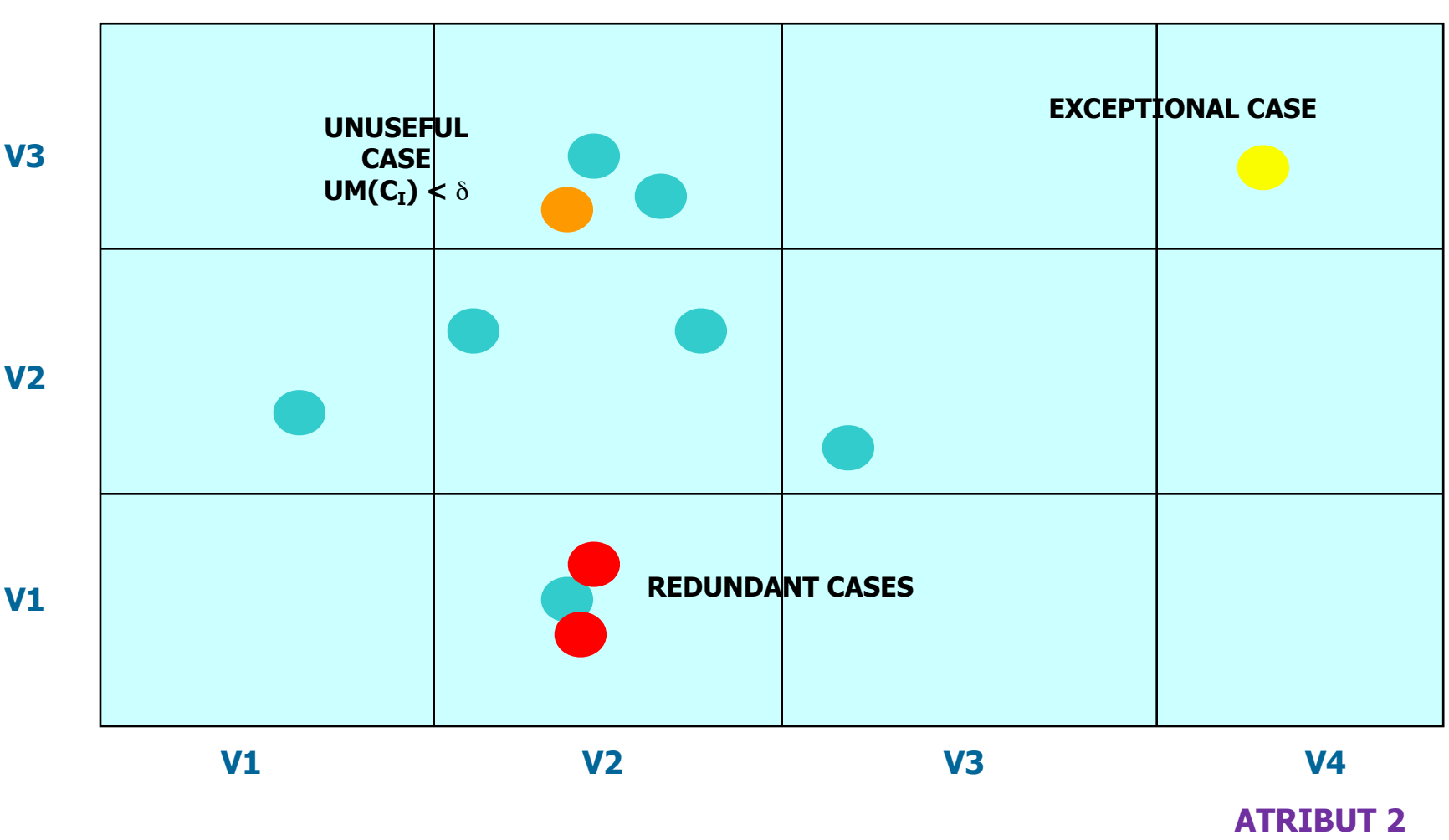
Espai: Quan aprendre/Oblidar casos?

- **Categorització de casos**
- **Aprendre només casos rellevants**
  - **Mesura de rellevància**
- **Oblidar casos inútils** i casos **no excepcionals**
  - **Mesura d'utilitat**

Ontologia de casos (1)

- Useful case (C_i) $\Leftrightarrow$ UM(C_i) $\geq \delta$
- Useless case (C_i) $\Leftrightarrow$ UM(C_i) $< \delta$
- Relevant case (C_i) $\Leftrightarrow$ Minimum $\{d(C_i,C_k)\} \geq \gamma$, where C_k are the cases in the same leaf of the case library tree than C_i, $k \neq i$
- Redundant case (C_i) $\Leftrightarrow$ Minimum $\{d(C_i,C_k)\} < \gamma$, where C_k are the cases in the same leaf of the case library tree than C_i, $k \neq i$
- Exceptional case (C_i) $\Leftrightarrow$ $\#(C_k) = 0$, where $\#(C_k)$ is the number of cases in the same leaf of the case library tree than C_i, $k \neq i$
- Normal case (C_i) $\Leftrightarrow$ $\#(C_k) > 0$, where $\#(C_k)$ is the number of cases in the same leaf of the case library tree than C_i, $k \neq i$

Ontologia de casos (1)

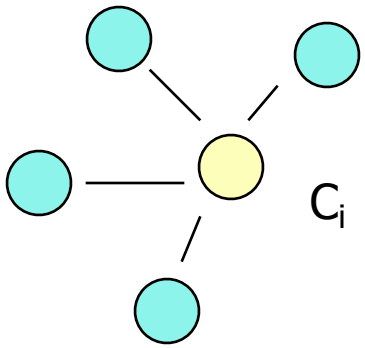


Aprenentatge: mesura de la rellevància

Quan **aprendre un cas nou** ?

Mesura de rellevàncica basada en mesura de similitud:

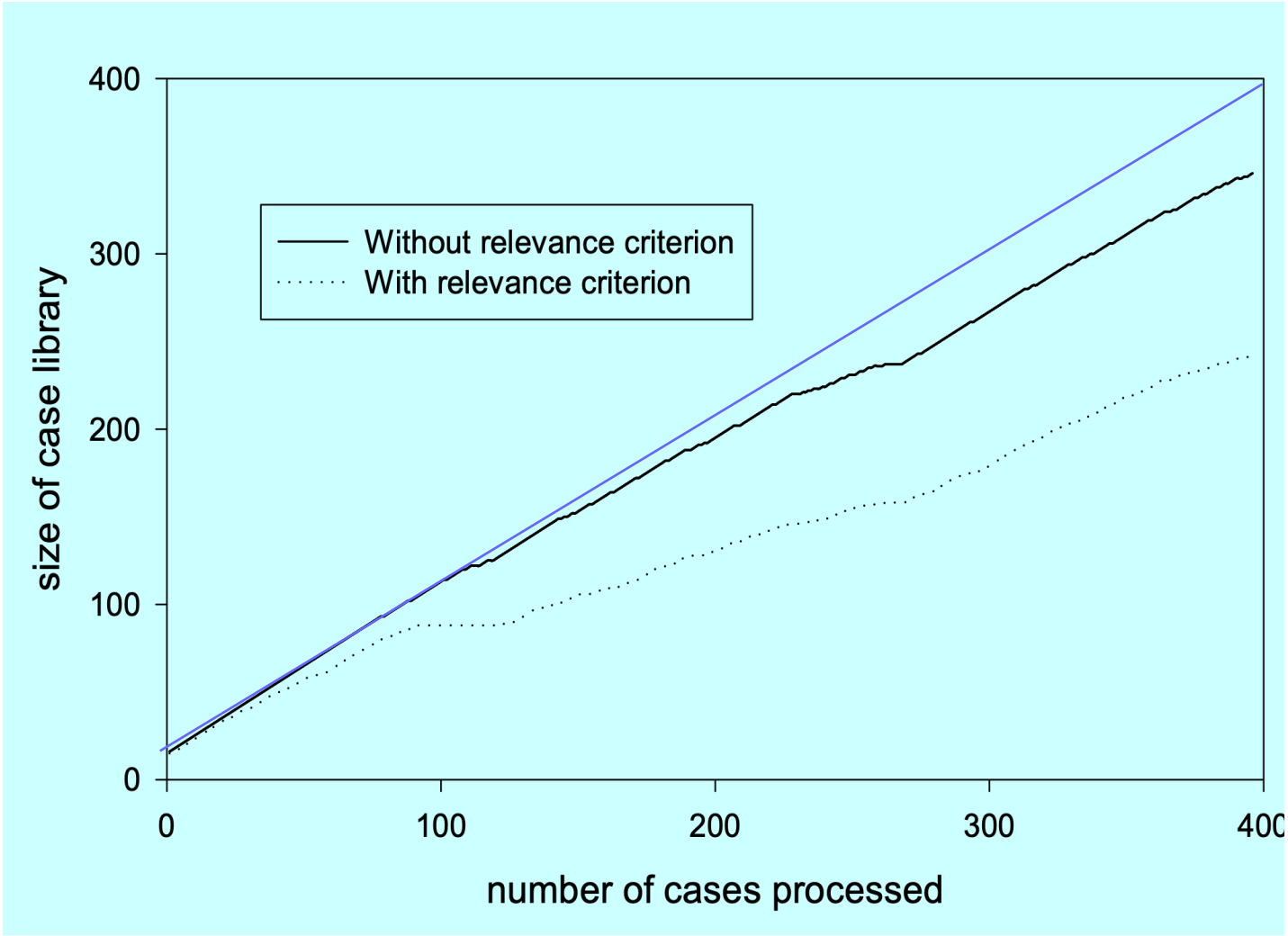
Aprendre un nou cas (C_i) $\Leftrightarrow$ Mínim $\{d(C_i,C_l)\} \geq \gamma$



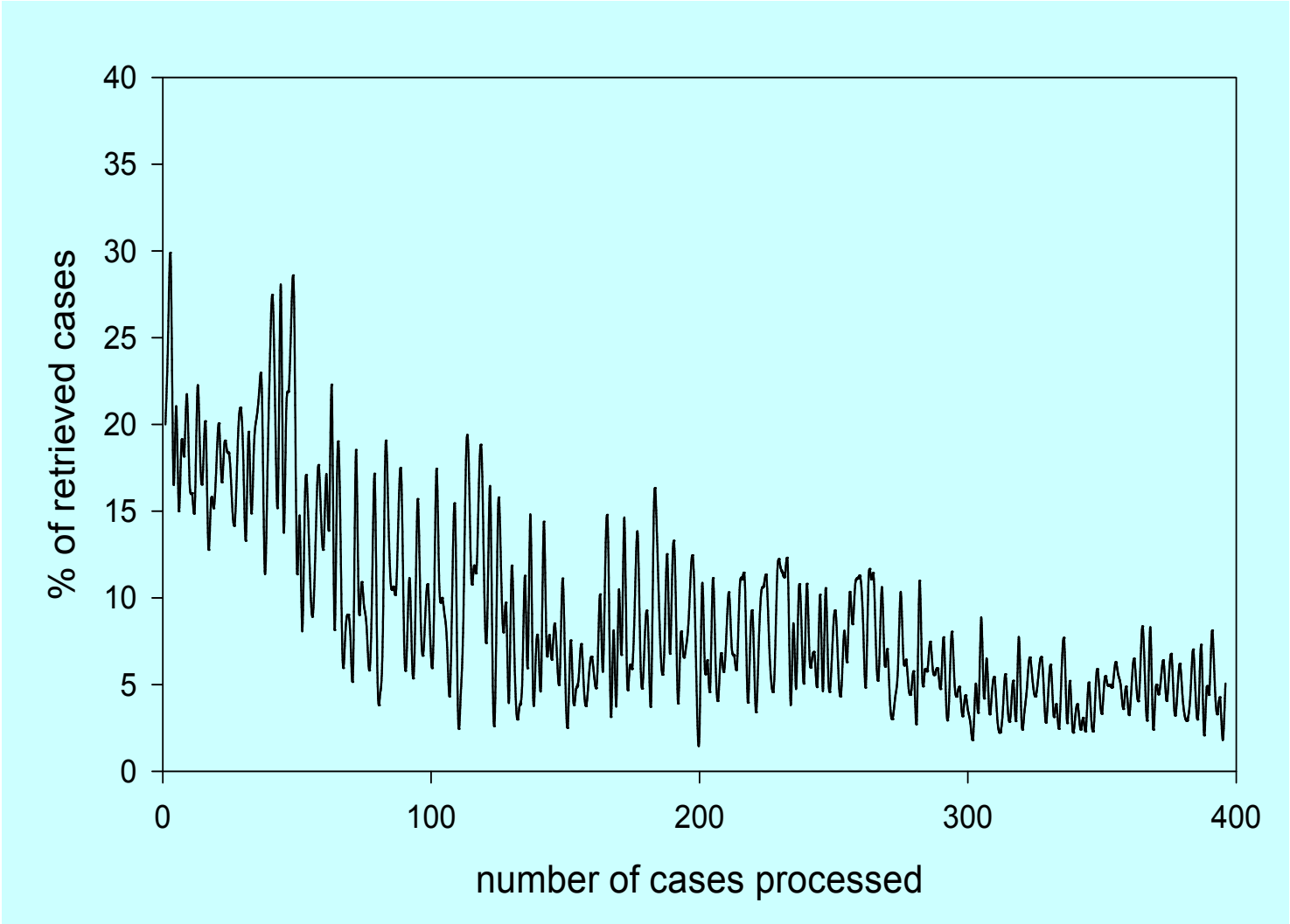
Comprovació experimental 1 (1)

- **Llibreria de casos inicial**: 15 casos representatius del conjunt de dades del 96/97 de l'EDAR de Girona
- **Experiment 1**
  - Aprendre el cas ⇔ **#casos-mateixa-fulla ≤ 3**
  - 396 casos processats, 87.3% apresos
- **Experiment 2**
  - Aprendre només els casos **rellevants**
  - 396 casos processats, 61% apresos

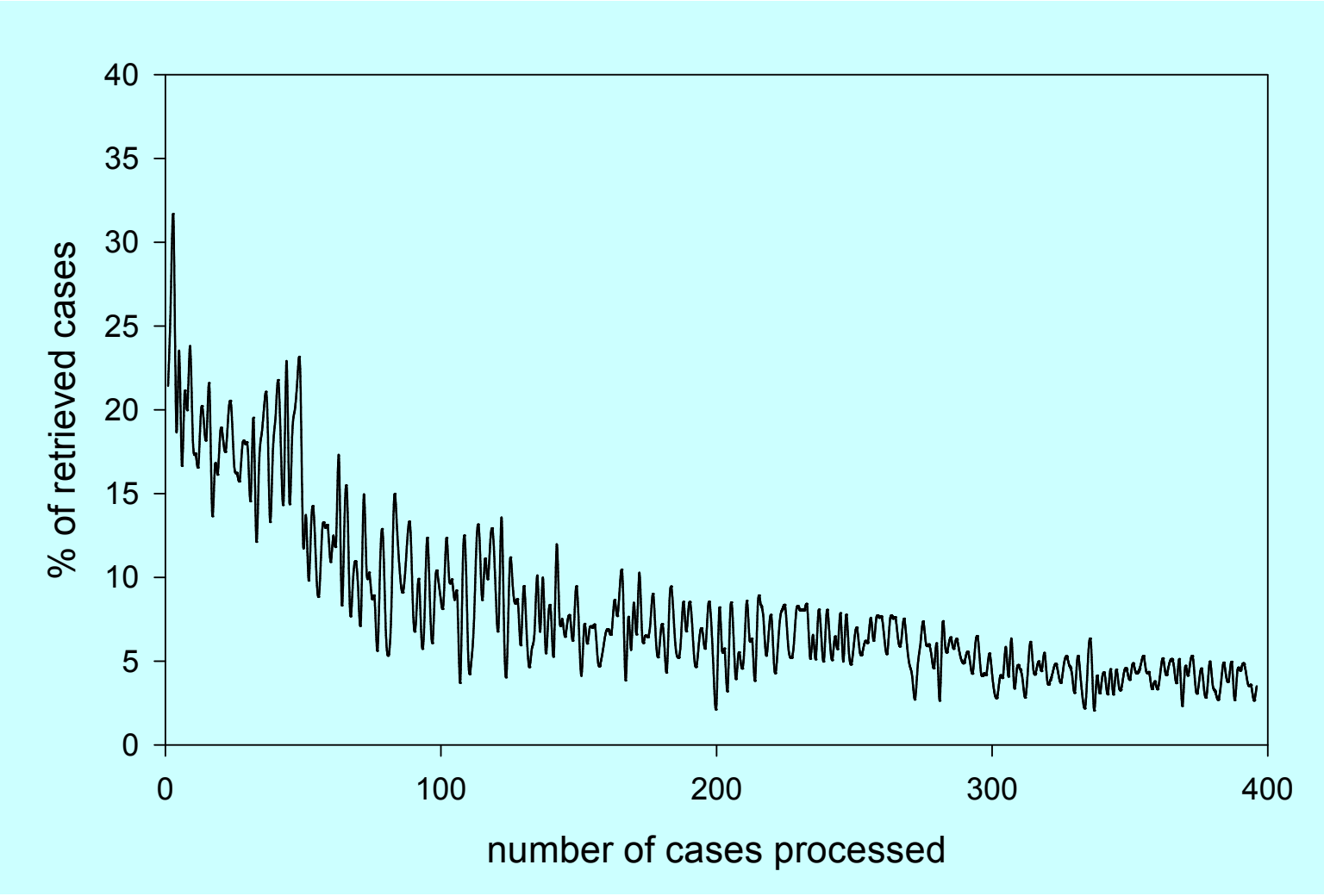
Comprovació experimental 1 (2)



Comprovació experimental 1 (3)



Comprovació experimental 1 (4)



Comprovació experimental 2 (1)

[Comas et al., 2001]

- 243 **objectes** (dia de funcionament de l'EDAR de Lloret)
- 63 **atributs** (quantitatius i qualitatius)
- **Atribut de classe**: l'estat de funcionament de la planta
- **Valors que falten ("missing")**
- 20 **classes** obtigudes d'un **procés de clustering**

Comprovació experimental 2 (2)

Situation	Class #	Nº of days
Normal WWTP-operation in winter days	1	81
Normal WWTP-operation in summer days	2	55
Rainy days	3	3
Storm days	4	3
Underloading	5	12
Overloading	6	1
Nitrification	7	2
Deflocculation	8	5
Bulking sludge due to Thiotrix (affecting the effluent)	9	3
Foaming sludge due to *Microthrix* with *normal* microfauna biodiversity	10	17
Summer days with optimal WWTP-operation	11	24
Clhorine shock	12	1
Denitrification in the secondary settler (*rising*)	13	7
Transition to a bulking-sludge episode due to Thiotrix	14	2
Weak episode of Foaming sludge due to *Nocardia*	15	4
Severe episode of foaming sludge due to *Nocardia*	16	5
Foaming sludge due to *Nocardia* and defflocculation	17	8
Foaming sludge due to *Microthrix* with *very low* microfauna *biodiversity*	18	1
Foaming sludge due to *Microthrix* and viscous bulking due to *Zooglea*	19	6
Winter-summer Plant configuration change	20	3

Comprovació experimental 2 (3)

- **10-fold stratified cross validation**
  - Conjunt sencer de 243 exemples
  - 10 conjunts de proves de 24/25 exemples
  - 10 conjunts d'entrenament de 219/218 exemples
- **Característiques observades**
  - Precisió predictiva al conjunt de proves
  - Precisió predictiva en tot el conjunt
  - Nombre d'exemples utilitzats
  - Nombre de funcions utilitzades
  - Interpretació significativa

Comprovació experimental 2 (4)

Nº Attrib.	Type of library	Case Retrieval Accuracy (%)		
		First	Second	Predominant
19	Plain memory	65.8	59.7	68.7
	Hierarchical (relevant cases)	62.5	44.9	52.3
	Hierarchical (all cases)	64.2	44.4	51.1
63	Plain memory	68.7	60.5	70.4

Method	Number of Attributes	Number of Examples	Prediction accuracy on test set (%)	Meaningful Interpretation	Prediction accuracy on whole data set (%)
C4.5 (63 atts)	24	243	63.51	Partially	89.7
CN2 (63 atts)	44	243	63.98	Partially	98.8
BPRI (63 atts)	63	243	58.9	Partially	-
k-NN (63 atts)	63	243	76.38	No	100
J48 (63 atts)	-	243	64.4	Partially	-
J48, bagging with 10 iterations	-	243	70.7	No	-
48, AdaBoostM1 with 10 iterations	-	243	73.6	No	-
C4.5 (19 atts)	11	243	65.11	Mostly	87.2
CN2 (19 atts)	19	243	65.45	Mostly	95.9
k-NN (19 atts)	19	243	71.22	No	100
Opencase (plain memory)	19	243	68.73	No	100
encase (hierarchical, relevant cases)	19	220	62.50	Yes	97.1
Opencase (hierarchical, all cases)	19	243	64.20	Yes	98.8
Opencase (plain memory)	63	243	70.40	No	100

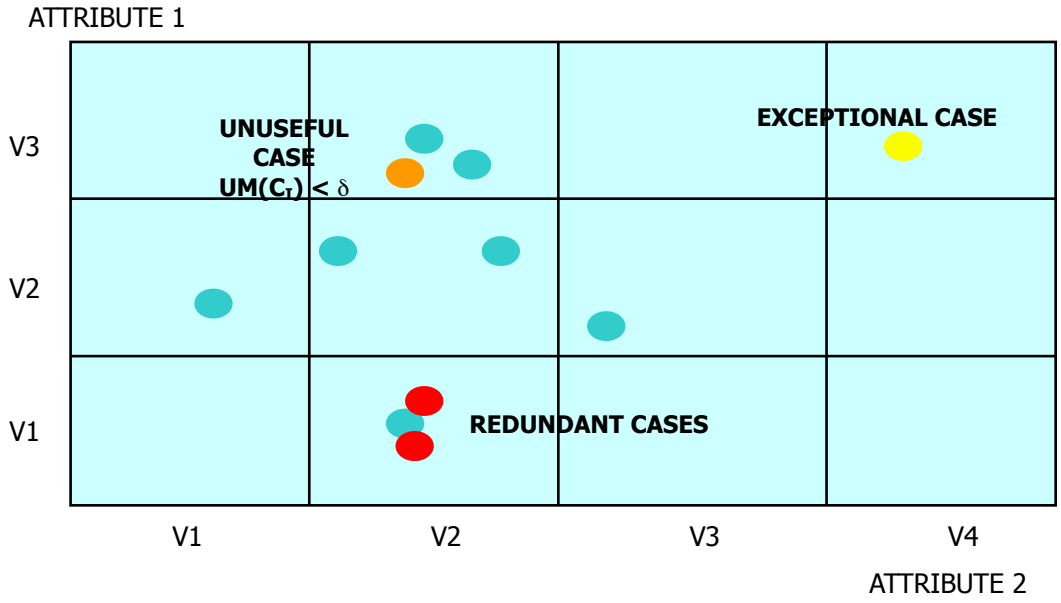
Espai
Quan oblidar?
[Sànchez-Marrè et al., 1997]

Oblit

Quan cal **oblidar un cas** ?

El criteri d'oblit està basat en una **mesura d'utilitat** i en la **categorització dels casos**:

Oblidar el cas (C_i) $\Leftrightarrow$ Unuseful(C_i) and Normal(C_i)



Competència

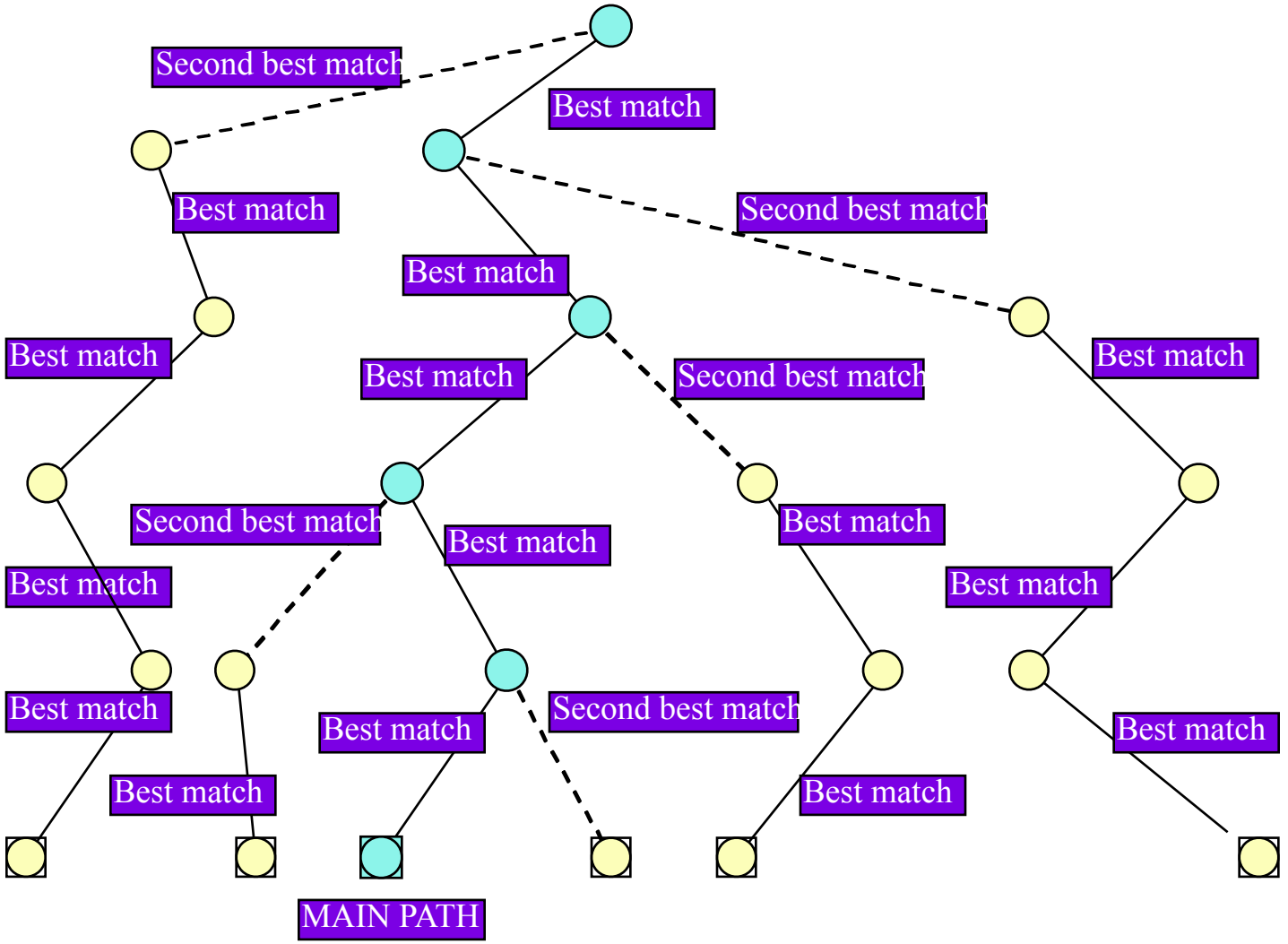
Cerca infructuosa / pèrdua de casos òptims

Cerca infructuosa a la Biblioteca de Casos

- **Re-exploració** [Sànchez-Marrè et al., 1997]
- **Meta-casos** [Sànchez-Marrè et al., 2000]

Re-exploració

[Sànchez-Marrè et al., 1997]

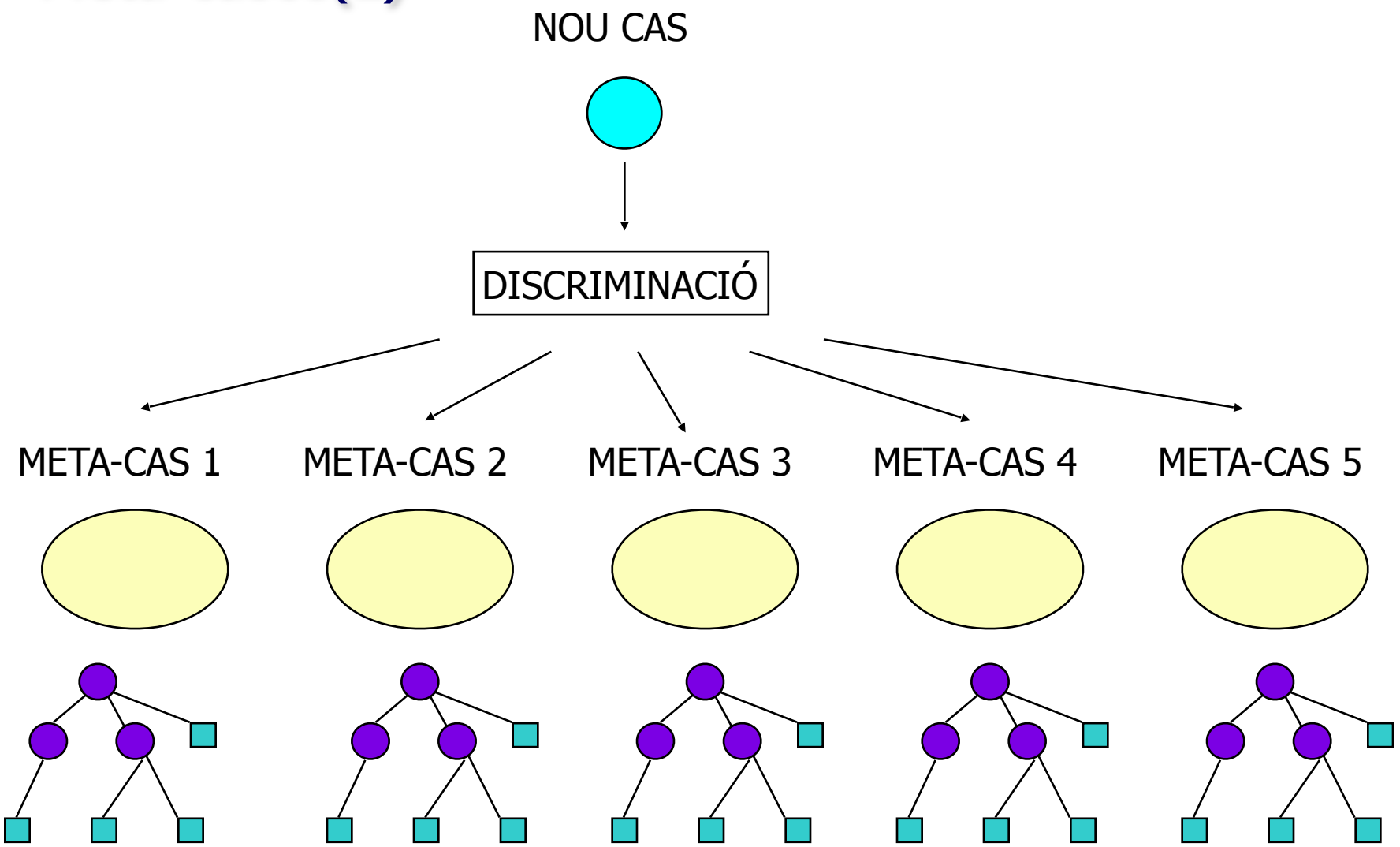


Meta-casos (1)

[Sànchez-Marrè et al., 2000]

- **Definició d'un conjunt de casos prototípics**: els **meta-casos**
- **Induir el conjunt relacionat de biblioteques de casos**
  - Diferents característiques rellevants
  - Ordenació discriminant diferent
  - Ponderació diferent
- **Recuperar de la/les biblioteca/biblioteques de casos més rellevants**

Meta-casos(2)



Comprovació Experimental (1)

- **Biblioteca de casos inicial**: 15 casos representatius del conjunt de dades 96/97 de l'EDAR de Girona
- **Experiment 1**
  - Utilitzant la biblioteca de casos estàndard
- **Experiment 2**
  - Utilitzant els metacasos i el conjunt de 5 biblioteques de casos definides

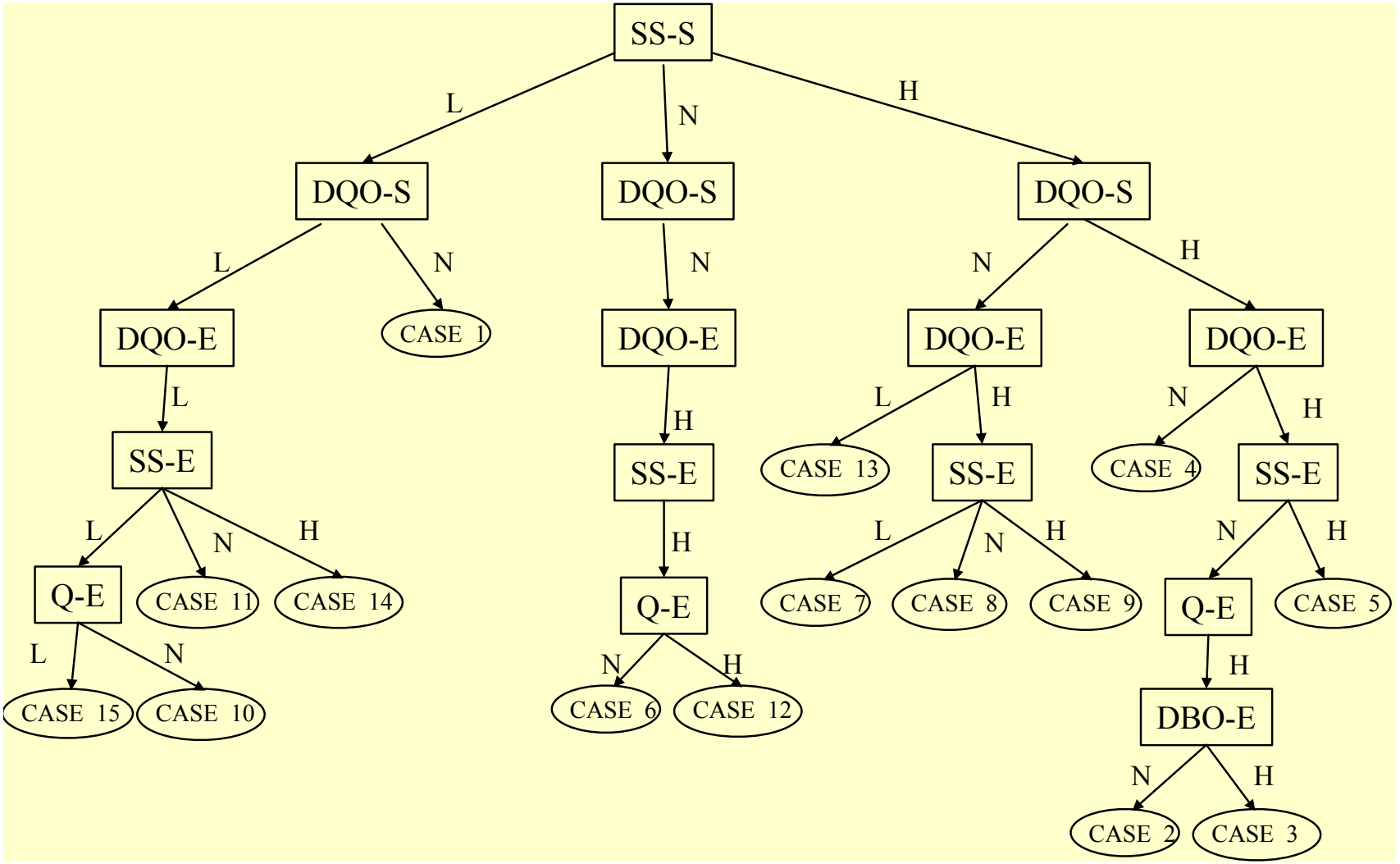
Comprovació Experimental (2)

			Modalities		
Feature	**Interpretation (units)**	**Weight**	**Low**	**Normal**	**High**
SS-S	Suspended solids at the output of the plant (mg/l)	9	(0 – 10)	(10 – 20)	(20 – 100)
DQO-S	Chemical oxidizable organic matter at the output (mg/l)	9	(0 – 30)	(30 – 70)	(70 – 200)
DQO-E	Chemical oxidizable organic matter at the input (mg/l)	7	(0 – 300)	(300 – 500)	(500 – 1000)
SS-E	Suspended solids at the input of the plant (mg/l)	6	(0 – 150)	(150 – 300)	(300 – 750)
Q-E	Inflow wastewater (m^3/d)	5	(0 – 30000)	(30000 – 40000)	(40000 – 60000)
DBO-E	Biodegradable organic matter at the input (mg/l)	8	(0 – 100)	(100 – 250)	(250 – 600)
DQO-D	Chemical oxidizable organic matter at the output of the primary settler(mg/l)	6	(0 – 150)	(150 – 300)	(300 – 600)
SS-D	Suspended solids at the output of the primary settler (mg/l)	6	(0 – 80)	(80 – 200)	(200 – 450)
IVF	Sludge volume index (ml)	8	(0 – 125)	(125 – 220)	(220 – 400)
V30	Measure of the sedimentability of the activated sludge (ml/g)	5	(0 – 150)	(150 – 250)	(250 – 450)

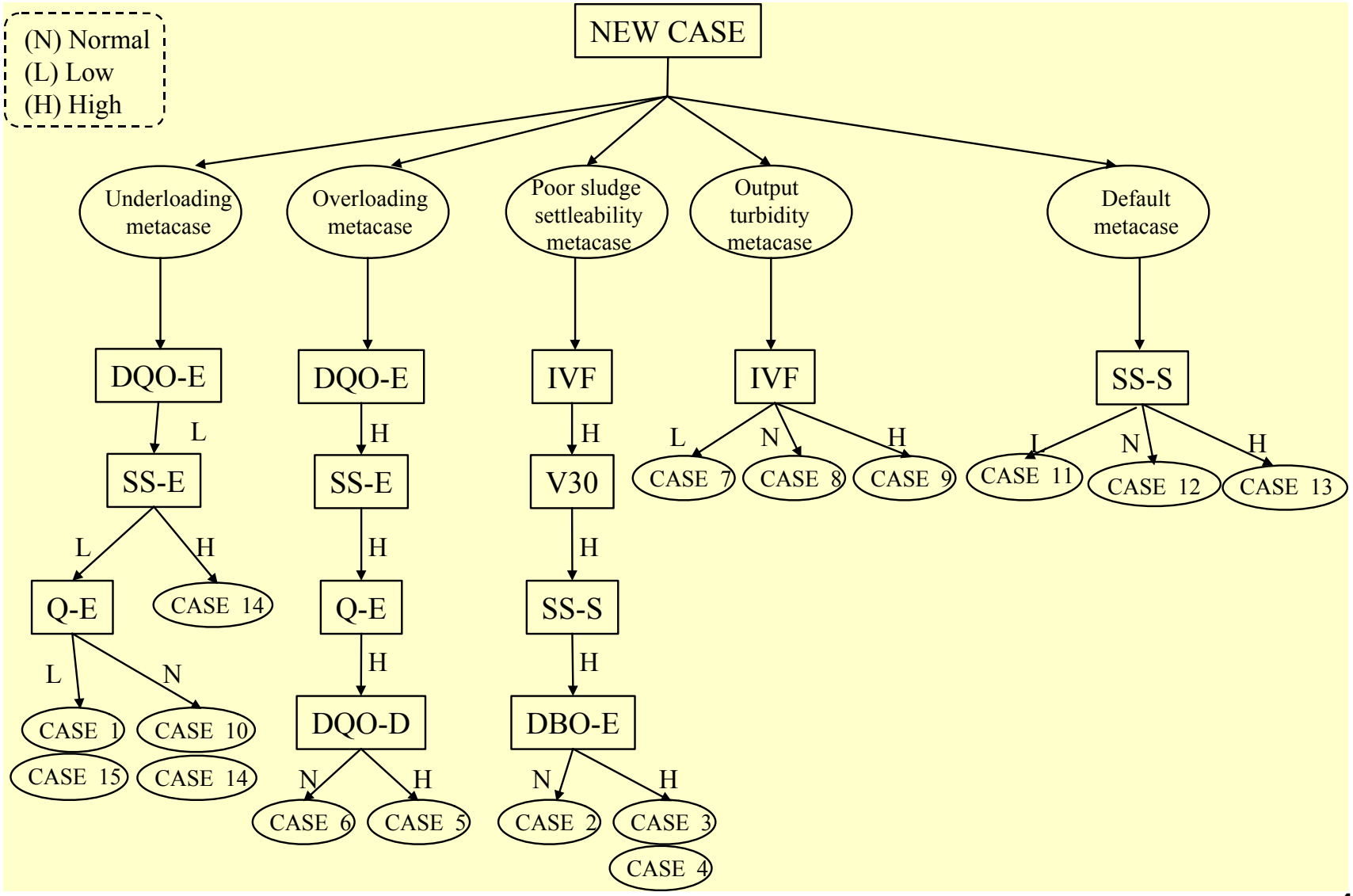
Comprovació Experimental (3)

	Libraries set				
Discriminant order	**Standard Library**	**Underloading**	**Overloading**	**Poor sludge settleability**	**Turbidity**
1	SS-S	DQO-E	DQO-E	IVF	IVF
2	DQO-S	SS-E	SS-E	V30	SS-S
3	DQO-E	Q-E	Q-E	SS-S	DBO-E
4	SS-E		DQO-D	DBO-E	
5	Q-E		DBO-E		
6	DBO-E				
7	DQO-D				
8	SS-D				
9	IVF				
10	V30				

Aproximació estàndard



Aproximació amb Meta-Casos



Resultats de la recuperació de casos

Optimal case retrieval	Standard Retrieval (1 library)	Meta-cases retrieval (Libraries set)					
		Standard	Underloading	Overloading	Poor sludge settleability	Turbidity	Total
First (%)	21	75	71	100	29	75	68
Other (%)	54	100	86	100	100	100	93

Referències (1)

- [Comas et al., 2001] J. Comas, S. Dzeroski, K. Gibert, I. R.-Roda and M. Sànchez-Marrè (2001). Knowledge discovery by means of inductive methods in wastewater treatment plant data. AI Communications 14(1):45-62. January 2001. ISSN 0921-7126.
- [Sànchez-Marrè, 1996] M. Sànchez-Marrè. DAI-DEPUR: an integrated supervisory multi-level architecture for wastewater treatment plants. Ph. D. Thesis. Dept. de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya. 1996.
- [Sànchez-Marrè et al., 2000] M. Sànchez-Marrè, U. Cortés, I. R.-Roda & M. Poch (2000). Using Meta-cases to Improve Accuracy in Hierarchical Case Retrieval. Computación y Sistemas 4(1):53-63, July 2000.
- [Sànchez-Marrè et al., 1997] Sànchez-Marrè, M., Cortés, U., R-Roda, I., Poch, M. & Lafuente, J. (1997). Learning and Adaptation in Wastewater Treatment Plants through Case–Based Reasoning. Computer-Aided Civil and Infrastructure Engineering 12(4):251-266.

Intelligent Data Science and Artificial Intelligence (IDEAI-UPC)

Miquel Sànchez-Marrè
miquel@cs.upc.edu

Knowledge Engineering and Machine Learning Group
UNIVERSITAT POLITÈCNICA DE CATALUNYA

https://kemlg.upc.edu