Clustering (IA)

Classificació conceptual (Michalski & Stepp 1983)

Lògica formal (instància vs objecte mostral)

Variables (atributs) qualitatius

Input: K nro de classes

- Triar K instàncies (random) com a llavor de cada classe
  - Randomly
  - Strategically
- Associar a cada classe el concepte induït de cada llavor

– (A)ADetectar quin és l'atribut a generalitzar de cada classe tal que
  - Contingui el màxim nro d'objectes observats
  - Minimitzi l'sparsness
  - S'evitin sol.lapaments entre classes

– Avaluar els nous conceptes de classe sobre la matriu d'instàncies i recollir els que satisfan cada concepte

– Ajustar els conceptes de classe estrictament a les instàncies identificades en el pas anterior

– Repetir des d'A fins que totes les instàncies es col.loquin en alguna classe

Clustering (IA)

Conceptual clustering (Michalski & Stepp 1983)

	Age	Weight	cigarrettes	Hard attacks
	years	*Kg*	*Pack/week*	#
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2

Choose 2 random seeds: M, R

Associate concepts:

M: [Age=30]&[Weight=Medium]&[Cigarrettes=Low]&[H.Atack=0]

R: [Age=30]&[Weight=Low] &[Cigarrettes=Moderate]&[H.Atack=0]

Conceptual clustering (Michalski & Stepp 1983)

Goal: Find K=2 classes

Initial random seeds:

M: [Age=30]&[Weight=Medium]&[Cigarrettes=Low]&[H.Atack=0]

R: [Age=30]&[Weight=Low] &[Cigarrettes=Moderate]&[H.Atack=0]

	Age	Weight	cigarrettes	Hard attacks
	years	*Kg*	*Pack/week*	#
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2

Conceptual clustering (Michalski & Stepp 1983)

Goal: Find K=2 classes

Initial random seeds:

M: [Age=30]&[Weight=Medium]&[Cigarrettes=Low]&[H.Atack=0]

R: [Age=30]&[Weight=Low] &[Cigarrettes=Moderate]&[H.Atack=0]

Generalizations: variables with differences in M and R: Weight, Cigarrettes

1) Grow M: weight ≠Low

Reevaluate and find class concept

M1: [Weight=Medium, High]

	Age	Weight	cigarrettes	Hard attacks
	years	*Kg*	*Pack/week*	#
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2

Conceptual clustering (Michalski & Stepp 1983)

Goal: Find K=2 classes

Initial random seeds:

M: [Age=30]&[Weight=Medium]&[Cigarrettes=Low]&[H.Atack=0]

R: [Age=30]&[Weight=Low] &[Cigarrettes=Moderate]&[H.Atack=0]

Generalizations: variables with differences in M and R: Weight, Cigarrettes

1) Grow M: weight ≠Low

M1: [Weight=Medium, High]

2) Grow R: weight ≠ Medium

R2: [weight=Low, High]

	Age	Weight	cigarrettes	Hard attacks
	years	*Kg*	*Pack/week*	#
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2

Conceptual clustering (Michalski & Stepp 1983)

Goal: Find K=2 classes

Initial random seeds:

M: [Age=30]&[Weight=Medium]&[Cigarrettes=Low]&[H.Atack=0]

R: [Age=30]&[Weight=Low] &[Cigarrettes=Moderate]&[H.Atack=0]

Generalizations: variables with differences in M and R: Weight, Cigarrettes

1) Grow M: weight ≠Low

M1: [Weight=Medium, High]

2) Grow R: weight ≠ Medium

R2: [weight=Low, High]

3) Grow M: Cigarrettes ≠ Moderate

M3: [Cigarrettes=Low, High]

	Age	Weight	cigarrettes	Hard attacks
	years	*Kg*	*Pack/week*	#
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2

Conceptual clustering (Michalski & Stepp 1983)

Goal: Find K=2 classes

Initial random seeds:

M: [Age=30]&[Weight=Medium]&[Cigarrettes=Low]&[H.Atack=0]

R: [Age=30]&[Weight=Low] &[Cigarrettes=Moderate]&[H.Atack=0]

Generalizations: variables with differences in M and R: Weight, Cigarrettes

1) Grow M: weight ≠Low

M1: [Weight=Medium, High]

2) Grow R: weight ≠ Medium

R2: [weight=Medium]&[H.Att ≠ 1]

3) Grow M: Cigarrettes ≠ Moderate

M3: [Weight=Low, High]

4) Grow R: Cigarrettes ≠ Low

R4: [Cigarrette=Moder, High]

	Age	Weight	cigarrettes	Hard attacks
	years	*Kg*	*Pack/week*	#
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2

Conceptual clustering (Michalski & Stepp 1983)

Goal: Find K=2 classes

Initial random seeds:

M: [Age=30]&[Weight=Medium]&[Cigarrettes=Low]&[H.Atack=0]

R: [Age=30]&[Weight=Low] &[Cigarrettes=Moderate]&[H.Atack=0]

Generalizations: variables with differences in M and R: Weight, Cigarrettes

A) Generalize M over R:

1) Grow M: weight ≠Low: M1: [Weight=Medium, High]

3) Grow M: Cigarrettes ≠ Moderate: M3: [Cigarrettes=Low, High]

B) Generalize R over M:

2) Grow R: weight ≠ Medium: R2: [weight=Low, High]

4) Grow R: Cigarrettes ≠ Low: R4: [Cigarrette=Moder, High]

This produces 4 possible descriptions of the 2-class partition:

P1={M1, R2}

P2={M1,R4}

P3={M3, R2}

P4={M3,R4}

Some overlapp!!!!!!! Specification!!!!!

Conceptual clustering (Michalski & Stepp 1983)

Goal: Find K=2 classes

Initial random seeds:

M: [Age=30]&[Weight=Medium]&[Cigarrettes=Low]&[H.Atack=0]

R: [Age=30]&[Weight=Low] &[Cigarrettes=Moderate]&[H.Atack=0]

Generalizations: variables with differences in M and R: Weight, Cigarrettes

P1= {M1, R2}

M1: [Weight=Medium, High]

R2: [Weight=Low, High]

Overlap!!!!! Specify R2,weight no high

Reevaluate and find class concept

M1: [Weight=Medium, High]

R1: [Age ≤40] &[Weight=Low]

Extension(P1)={M1={B,C,J,M,P,S,T}

R1={A,F,R}}

	Age	Weight	cigarrettes	Hard attacks
	years	*Kg*	*Pack/week*	#
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2

Conceptual clustering (Michalski & Stepp 1983)

Goal: Find K=2 classes

Initial random seeds:

M: [Age=30]&[Weight=Medium]&[Cigarrettes=Low]&[H.Atack=0]

R: [Age=30]&[Weight=Low] &[Cigarrettes=Moderate]&[H.Atack=0]

Generalizations: variables with differences in M and R: Weight, Cigarrettes

P1={M1,R1}

M1: [Weight=medium, high]

R1: [Age ≤40] &[Weight=Low]

P2= {M1, R4}

M1: [Weight=Low, High]

R4: [cigarrettes=Moderate, High]

Overlapping.

Specify R4&weight=Medium

And expand weight=Medium

	Age	Weight	cigarrettes	Hard attacks
	years	*Kg*	*Pack/week*	#
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2

Conceptual clustering (Michalski & Stepp 1983)

Goal: Find K=2 classes

Initial random seeds:

M: [Age=30]&[Weight=Medium]&[Cigarrettes=Low]&[H.Atack=0]

R: [Age=30]&[Weight=Low] &[Cigarrettes=Moderate]&[H.Atack=0]

Generalizations: variables with differences in M and R: Weight, Cigarrettes

P1={M1,R1}

M1: [Weight=medium, high]

R1: [Age ≤40] &[Weight=Low]

P2= {M1, R4}

M1: [Weight=Low, High]

R4: [cigarrettes=Moderate, High]

Overlapping.

Specify R4 and expand

Reevaluate concepts

M2: [Weight=Low, High]

R2: [Age ≤40]&[weight=Medium]

Extension(P2)={M2={C,MT}

R2={A,B,F,J,P,R,S}}

	Age	Weight	cigarrettes	Hard attacks
	years	*Kg*	*Pack/week*	#
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2

Conceptual clustering (Michalski & Stepp 1983)

Goal: Find K=2 classes

Initial random seeds:

M: [Age=30]&[Weight=Medium]&[Cigarrettes=Low]&[H.Atack=0]

R: [Age=30]&[Weight=Low] &[Cigarrettes=Moderate]&[H.Atack=0]

Generalizations: variables with differences in M and R: Weight, Cigarrettes

P3= M3, R2

M3=[Cigarretes=Low, High]

R2=[Weight=Low, High]

Overlapp, AND NO COVER!!!

R3:R2&[cigarrette=moderate]

expand [cig=mod]

Reevaluate concepts and expand

M3: [Cigarrettes=Low, High]

R3: [Age ≤40] &[Cigarrette=Moder]

Extension(P3)={M3={A,F,J,M,P,S,T}

R3={B,C,R}}

	Age	Weight	cigarrettes	Hard attacks
	years	*Kg*	*Pack/week*	#
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2

Conceptual clustering (Michalski & Stepp 1983)

Goal: Find K=2 classes

Initial random seeds:

M: [Age=30]&[Weight=Medium]&[Cigarrettes=Low]&[H.Atack=0]

R: [Age=30]&[Weight=Low] &[Cigarrettes=Moderate]&[H.Atack=0]

Generalizations: variables with differences in M and R: Weight, Cigarrettes

P3=

M3: [Weight=Low, High]

R3: [Age ≤40] &[Cigarrette=Moder]

P4= M3, R4

M3: [Cigarrette=Low, High]

R3: [Cigarrette=Moder,high]

Overlapping. Specify R3, cig no high

Reevaluate concepts

M4: [Age ≤40] & [Cigarrette=Low]& [H.Att≠1]

R4: [Cigarrette=Moder, High]

Extension(P4)={M4={F,J,M},
R4={A,B,C,P,R,S,T}}

	Age	Weight	cigarrettes	Hard attacks
	years	*Kg*	*Pack/week*	#
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2

Conceptual clustering (Michalski & Stepp 1983)

Goal: Find K=2 classes

Initial random seeds:

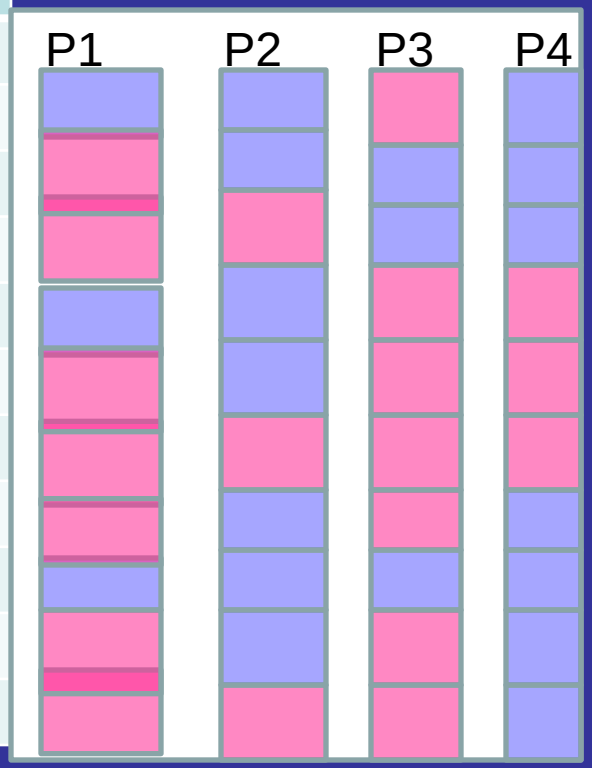
M: [Age=30]&[Weight=Medium]&[Cigarrettes=Low]&[H.Atack=0]

R: [Age=30]&[Weight=Low] &[Cigarrettes=Moderate]&[H.Atack=0]

Generalizations: variables with differences in M and R: Weight, Cigarrettes

Four possible generalizations:

	Age	Weight	cigarrettes	Hard attacks
	years	*Kg*	*Pack/week*	#
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2



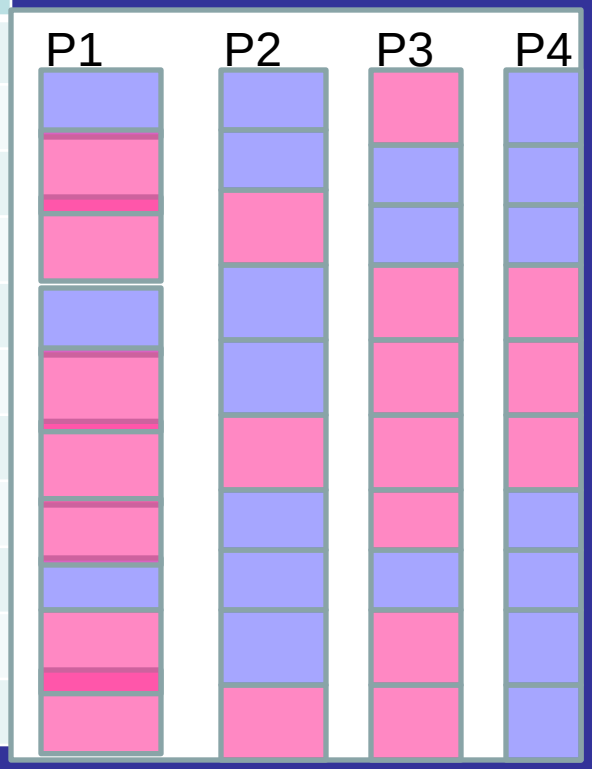
All directly understandable by user

Conceptual clustering (Michalski & Stepp 1983)

Goal: Find K=2 classes

Choose best generalization:

	Age	Weight	cigarettes	Hard attacks
	years	*Kg*	*Pack/week*	#
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2



Minimize over all clusterings P: $[\Sigma_{\alpha \epsilon P}\ r(\alpha)]/cardP$

Being $r(\alpha)=1-[p(\alpha)/t(\alpha)]$ the relative SPARSENESS of a concept α

t: total events covered by concept α

p: observed events covered by concept α

Conceptual clustering (Michalski & Stepp 1983)

SPARSENESS OF P1

P1={ M1: [Weight=Medium, High] ; R1: [Age ≤40] &[Weight=Low]}

Compute relative SPARSNESS of M1 and R1
Compute the MEAN RELATIVE SPARSNESS= [r(M1)+r(R1)]/2

t(M1): conceptos cubiertos por M1

#Age*#Weight*#Cigarr*#Heart.At
3 * 2 * 3 * 3 =54

p(M1): 7 observed cases

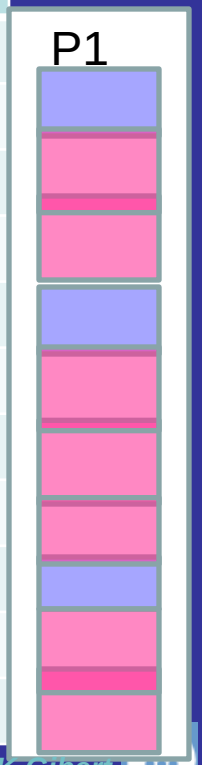
[47 non observed: (30, Med,high,0)]

r(M1)=1-7/47= 0.851

r(R1)= 1- 3/18 =0.83

r(P1)= (0.76+0.83)/2=0.795

	Age	Weight	cigarrettes	Hard attacks
	years	*Kg*	*Pack/week*	#
A	30	Low	High	1
B	40	High	Moderate	1
C	30	Medium	Moderate	2
F	40	Low	Low	2
J	30	High	Low	2
M	30	Medium	Low	0
P	40	High	High	0
R	30	Low	Moderate	0
S	50	High	High	2
T	40	Medium	High	2



Clustering (IA)

Conceptual Clustering (Michalski & Stepp 1983)
Formal logics (instances vs sample objects)
Qualitative Variables (atributes)

Input: K number of classes

- Choose K instances as class seeds
  - Randomly
  - Strategically

- Associate the concept induced from each seed to each class

– Detect the attributes for possible generalization
– Generate all possible generalizations, recalculate class concepts
– Correct overlappings by maximizing the coverage
– Select the proposal with minimal sparseness

– Respecify the winer to cover strictly the observed cases

– Repeat generalization till sparseness do not improve anymore

Final result: P={C1= {R,M,P,B,A}, C2={F,C,T,J,S}}