Automatic construction of concept hierarchies: the case of foliage-dwelling spiders

Martin Žnidaršič, Aleks Jakulin, Sašo Džeroski

Overview

- Data set characteristics
- Previous work
- Interaction analysis
- Constructive induction

Data set

- Domain: number of spider species in field margins
- Attributes:
  - Number of disturbance events per year
  - Field margin width, density
  - Herb cover
  - Slope direction
  - Many other life-form, soil and climate characteristics…
- 97 data instances
- Barthel and Plachter (1996), Anderlik-Wesinger et al. (1996)

Modelling

- Kampichler et al. (*Ecological Modelling* 2000)
  - Fuzzy rule-based model
  - Manually made hierarchy
  - Computer tuned rules
- Interaction analysis
  - Purely empirical taxonomy (hierarchy) of variables
- Constructive induction
  - Automatically generated hierarchy and rules (crisp)

Interaction analysis 1:

Attribute Dependencies

B \ A	Low diversity	High diversity	**Total margins**
Sparse margins	46%	25%	71%
Dense margins	3%	26%	29%
Total diversity	49%	51%	

Interaction assumption : model form P(A, B)

Independence assumption : model form P(A) P(B)

P(low_diversity, dense_margins) = 0.03

P(low_diversity) P(dense_margins) = 0.14

distinct deviation !

Interaction analysis 2:

2-way interaction

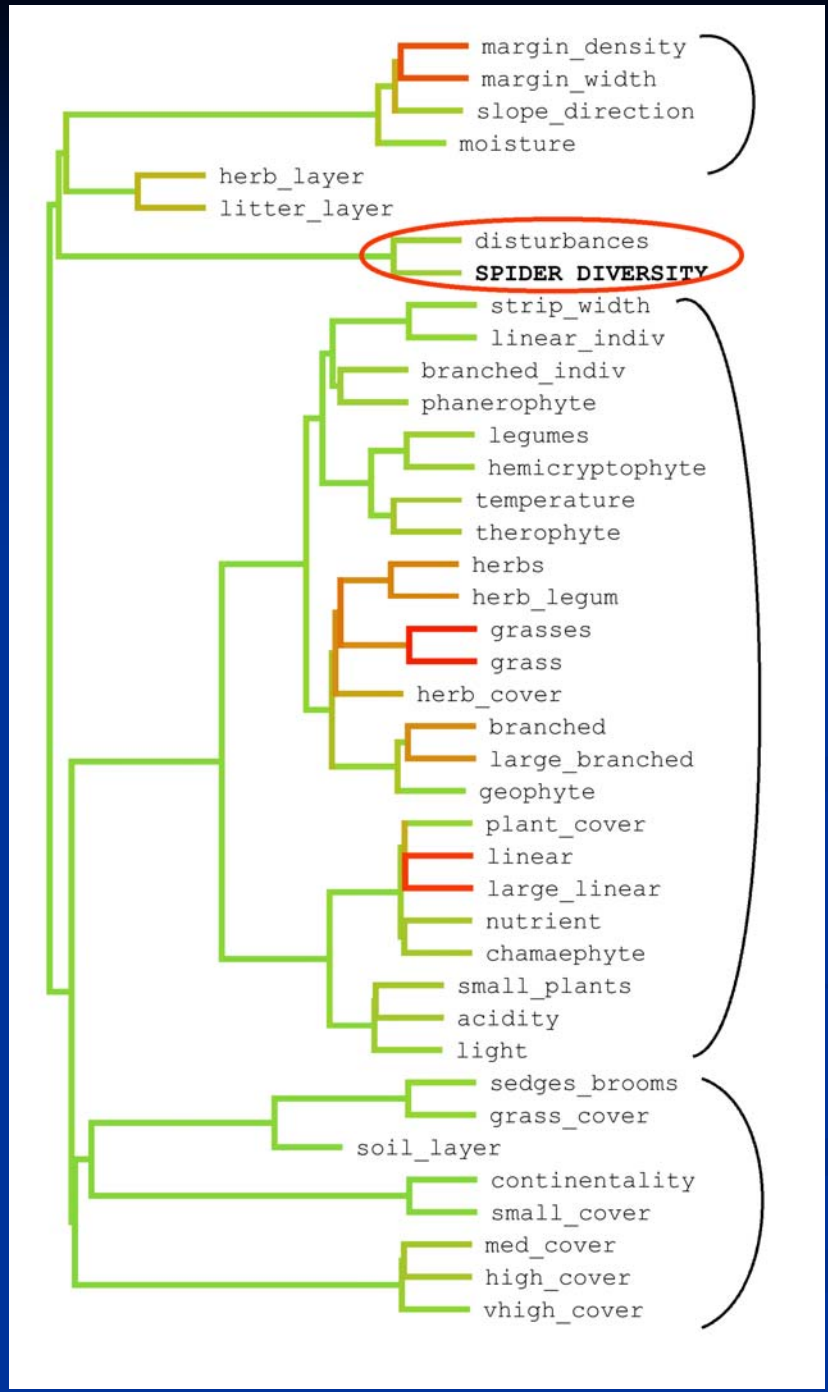
- Many measures of deviation
  - Kullback-Leibler divergence
  - Entropy

- Entropy:

$D(P(A,B) \,||\, P(A)\,P(B)) = H(A) + H(B) - H(A, B) = I(A; B)$

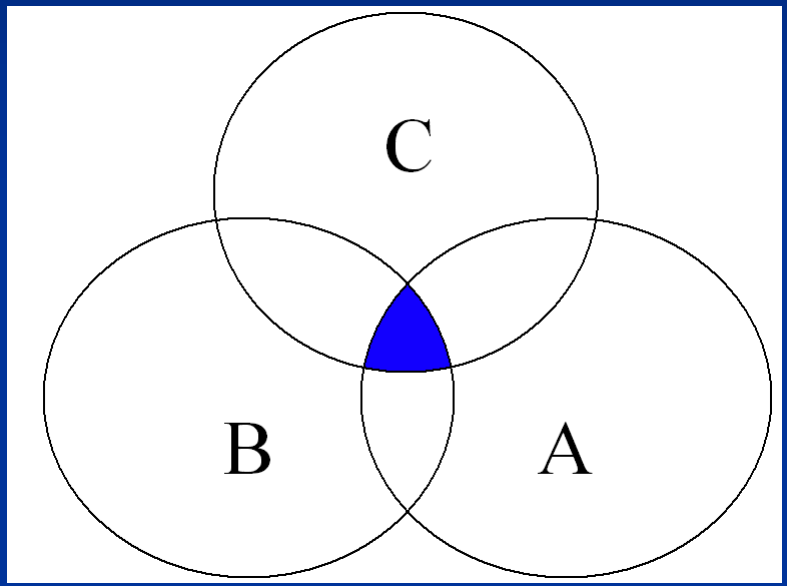
mutual information

- If mutual information is high, then A and B interact


margin_density
margin_width
slope_direction
moisture
herb_layer
litter_layer
disturbances
SPIDER DIVERSITY
strip_width
linear_indiv
branched_indiv
phanerophyte
legumes
hemicryptophyte
temperature
therophyte
herbs
herb_legum
grasses
grass
herb_cover
branched
large_branched
geophyte
plant_cover
linear
large_linear
nutrient
chamaephyte
small_plants
acidity
light
sedges_brooms
grass_cover
soil_layer
continentality
small_cover
med_cover
high_cover
vhigh_cover

Interaction analysis 4:

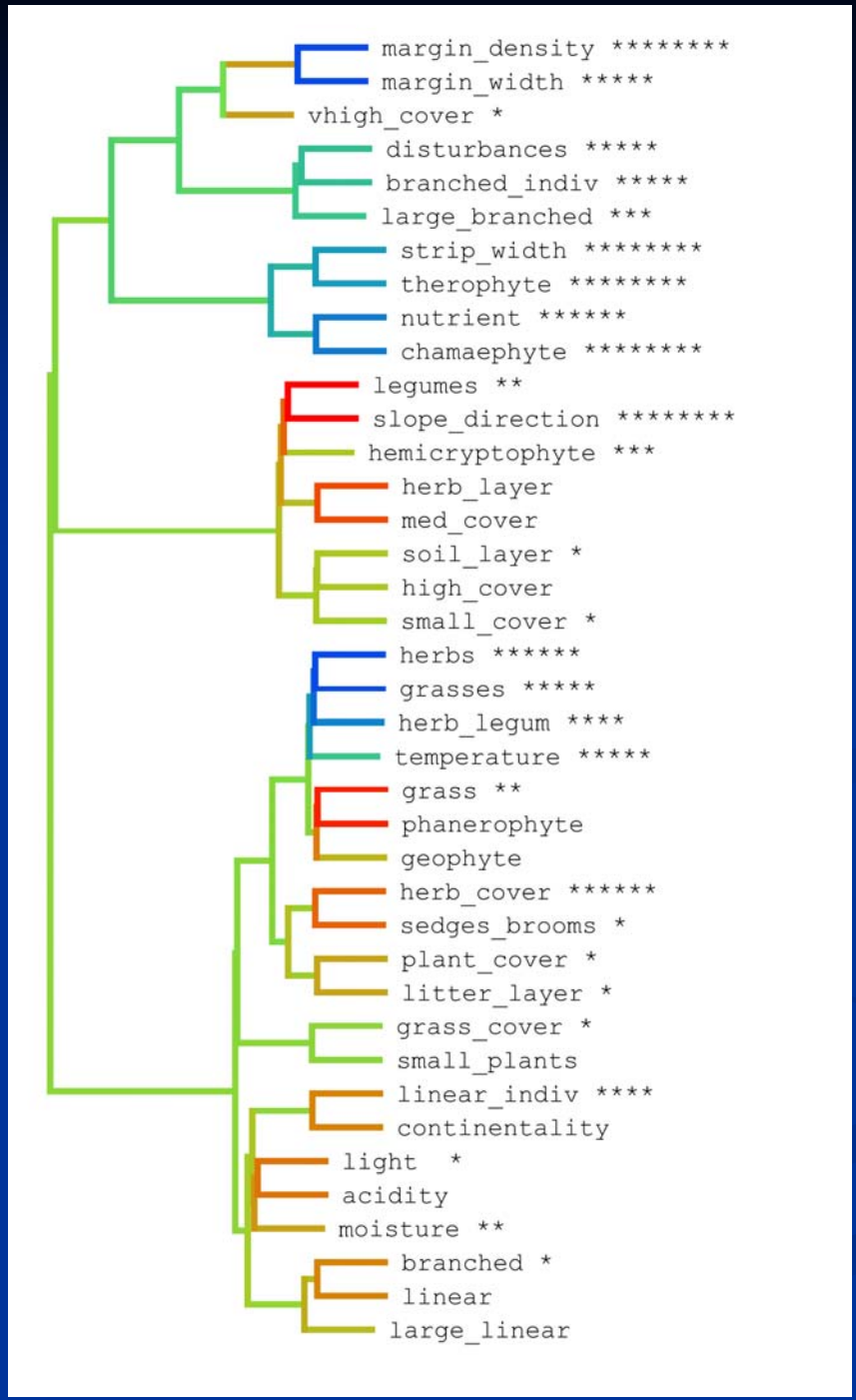
3-way interaction



$$I(A;B;C) := I(AB;C) - I(A;C) - I(B;C)$$

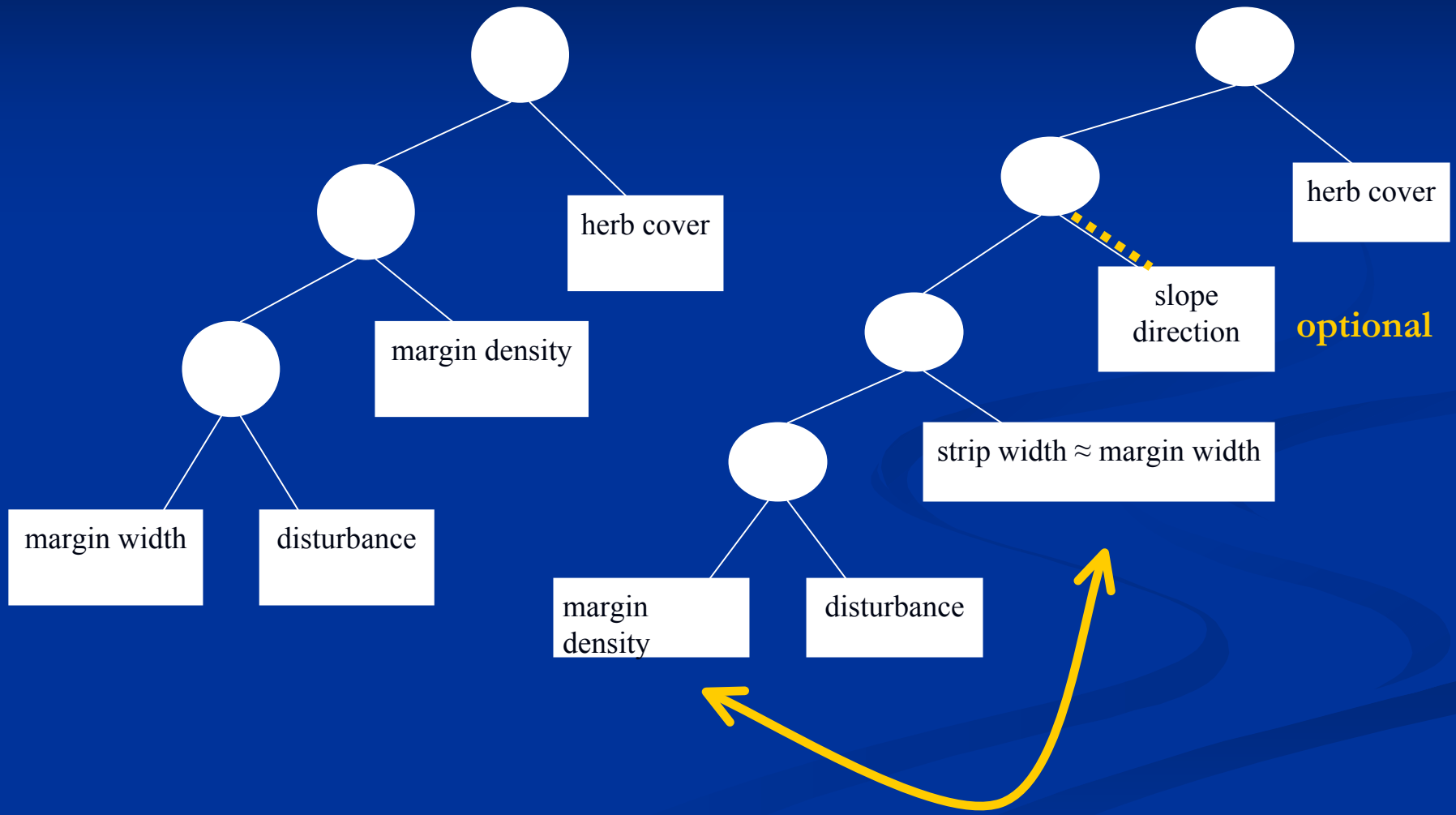
Positive I(A;B;C) means sinergy (more information together)

Negative I(A;B;C) means redundance


margin_density ********
margin_width *****
vhigh_cover *
disturbances *****
branched_indiv *****
large_branched ***
strip_width ********
therophyte ********
nutrient ******
chamaephyte ********
legumes **
slope_direction ********
hemicryptophyte ***
herb_layer
med_cover
soil_layer *
high_cover
small_cover *
herbs ******
grasses *****
herb_legum ****
temperature *****
grass **
phanerophyte
geophyte
herb_cover ******
sedges_brooms *
plant_cover *
litter_layer *
grass_cover *
small_plants
linear_indiv ****
continentality
light *
acidity
moisture **
branched *
linear
large_linear

Interaction analysis 6:

Comparison



Constructive induction

- Discovering concepts in data
- HINT – hierarchy induction tool
- Uses function decomposition
- Capable of:
  - Hierarchical structure construction
  - Creating new variables and rules

Constructive induction 2:

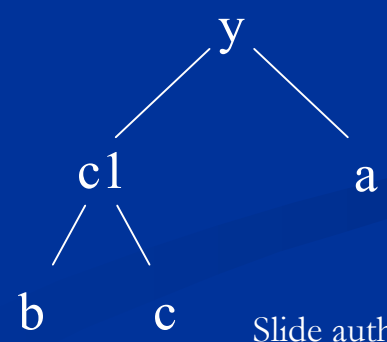
Function decomposition

a	b	c	y
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	1

a \ b c	0 0	0 1	1 0	1 1
0	0	0	0	1
1	1	1	1	1

b	c	c1
0	0	0
0	1	0
1	0	0
1	1	1

a	c1	y
0	0	0
0	1	1
1	0	1
1	1	1

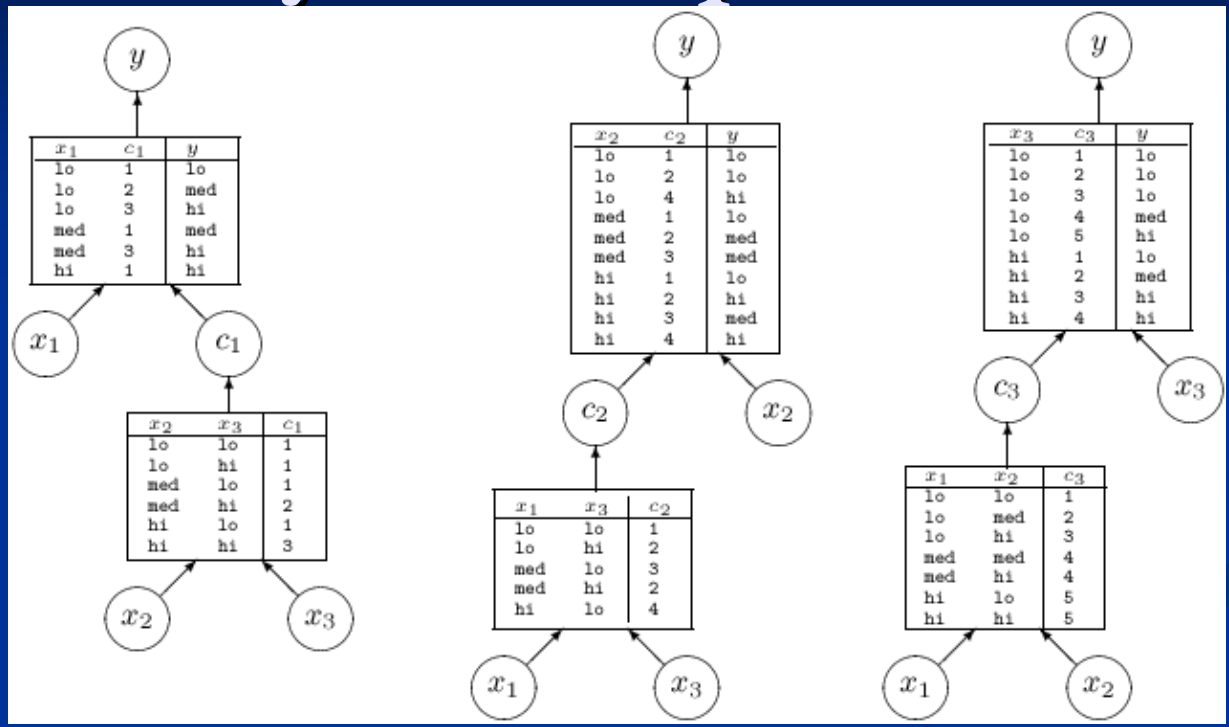


[y:= a ∨ b ∧ c]

Slide author: Janez Demšar

Constructive induction 3:

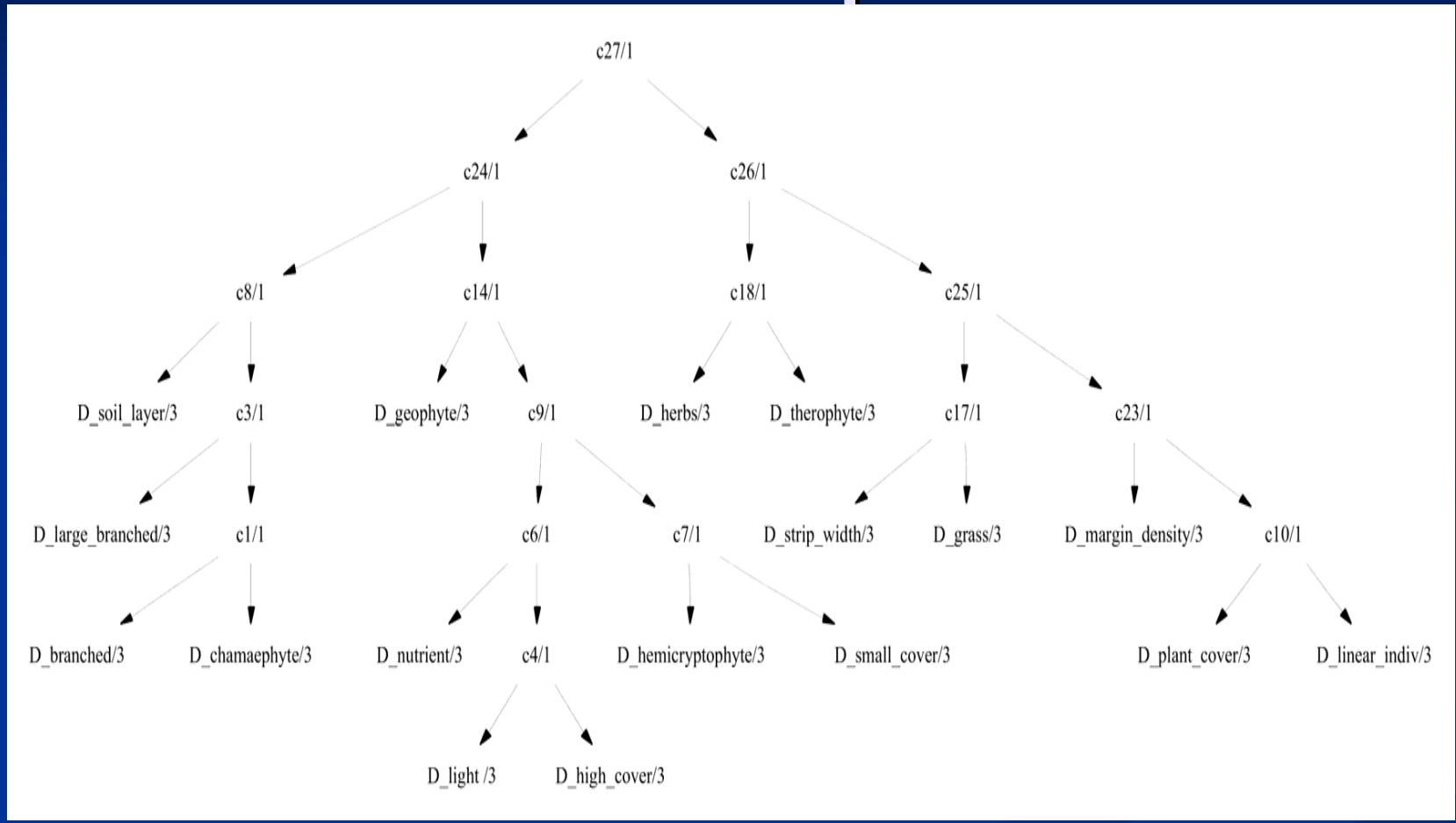
Many decompositions



- Which decomposition to select?
  - Smallest example set (rule set)
  - Smallest value set
  - Easiest interpretation

Constructive induction 4:

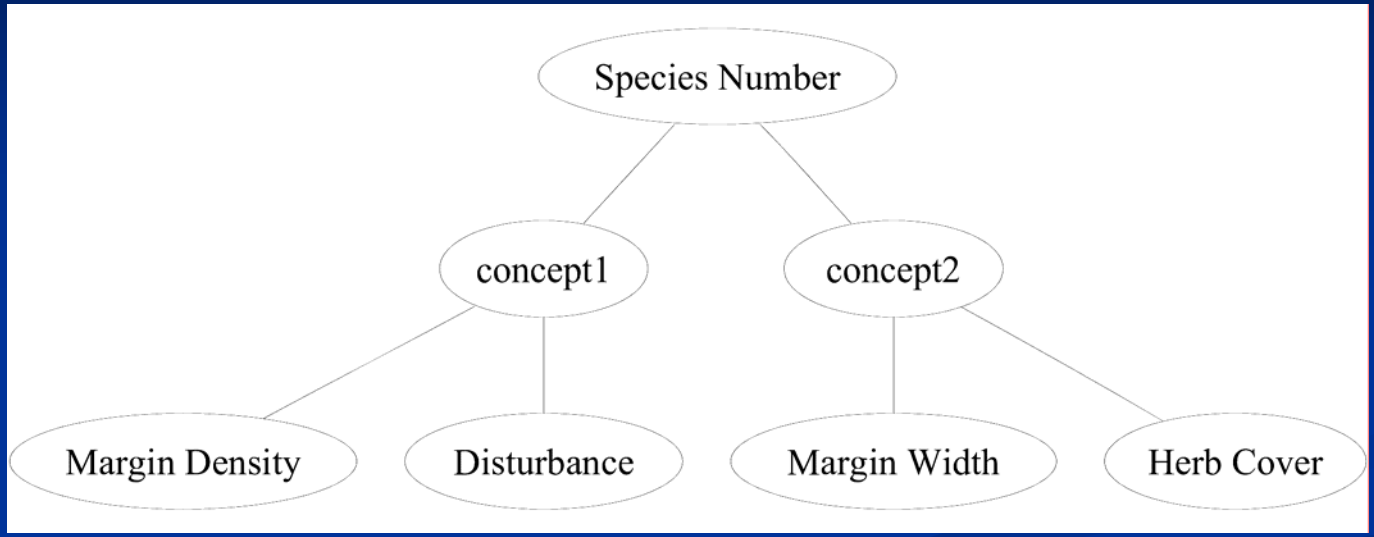
HINT on spiders



- No match in structure (of the 4 attributes)

Constructive induction 5:

Comparison (4 attributes)



- Only 4 atributtes, discretization as in Kampichler et al.
- Still, no match in structure
- But, direct comparison of performance is sound

Constructive induction 6:

Prediction performance

- Measure: mean absolute error
- Regression (Kampichler et al.) MAE : 3.17
- Fuzzy model (Kampichler et al.) MAE : 1.38
- HINT's model MAE : 2.69
- But our model is crisp!
- Crisp model (Kampichler et al.) MAE : 3.48
- **Advantage is fuzzy approach, not structure!**

Conclusion

- Two potentially useful methods
- Insight into variable relations with interaction analysis
- Complete initial model construction with HINT
- Another confirmation of advantages of fuzzy and probabilistic approaches → present work..