

# Interactivity Closes the Gap

Axel Blumenstock<sup>1,2</sup>, Jochen Hipp<sup>1</sup>, Steffen Kempe<sup>1</sup>,  
Carsten Lanquillon<sup>1</sup>, Rüdiger Wirth<sup>1</sup>

<sup>1</sup>DaimlerChrysler Research & Technology, Ulm

<sup>2</sup>University of Ulm, Dept. of Applied Information Processing

Data Mining for Business Applications  
Philadelphia, 2006-08-20

## About us

### Our group...

- ... has been engaged with data mining for years

### Our project customers...

- ... are experts in vehicle engineering
- ... must find explanations for quality issues quickly

### We...

- ... assess, enhance, and develop data mining techniques
- ... make them applicable for our users

## Task

To find explanations for specific quality issues with data mining, take...

- vehicles as instances
- label non-conforming vehicles positive
- vehicle descriptions as influence variables

...and any symbolic modeling technique will do?

## Task

To find explanations for specific quality issues with data mining, take...

- vehicles as instances
- label non-conforming vehicles positive
- vehicle descriptions as influence variables

...and any symbolic modeling technique will do?

➔ We experienced a gap between this theoretical idea and its practical application

# Data Properties

- **Mainly production and warranty data**
  - collected for operational, not analytical purposes
- **Uncertain class label**
  - multiple causes
  - counterpart of positives not truly negative
- **Highly imbalanced classes**
  - often way below 1 %
- **High-dimensional space**
  - 1000s of potential influence variables
- **Variables interact strongly**
  - non-causal variables show up as influences
- **True causes not in data**

# Approaches

## Classification

- separating the good from the bad
- “random” results due to strong “noise”

## Subgroup Discovery

- mining subsets with exceptional class distributions
- not sufficient for explanation

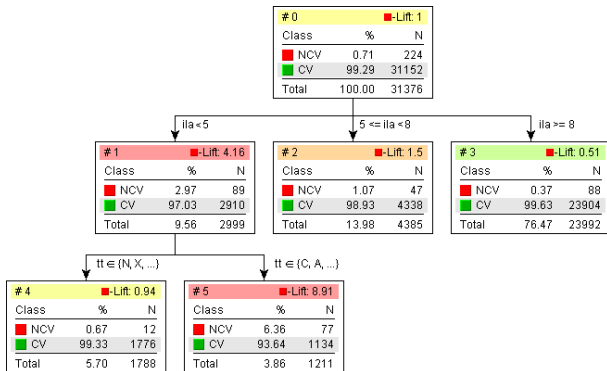
## Subgroup Description

- just does not omit the synonyms
- allows for knowledge acquisition *on the way*



# Decision Trees for Subgroup Discovery

- Search for subsets (paths) with high rates of positives



# Interactive Decision Trees

Variable	▲ wtLift	topRecall	topLift
● PDAT	0.300	75.53%	1.661
● Code_IA0	0.242	56.58%	1.747
● FZBM	0.222	28.16%	4.748
● Code_NZ1	0.196	39.74%	1.976
● NA	0.193	41.32%	1.878
● Rf	0.168	22.37%	4.019
● Code_IA5	0.154	20.79%	3.834
● NTI	0.138	19.21%	3.571
● Code_N55	0.134	57.89%	1.307

## Select split variable from a list

- Measure only needs to create a good ranking
- Very often, the expert-selected attribute is somewhat down the list

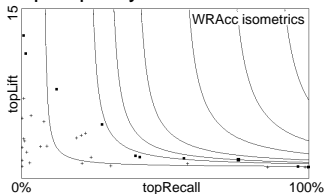
# Interactive Decision Trees

Variable	▲ wtLift	topRecall	topLift
● PDAT	0.300	75.53%	1.661
● Code_IA0	0.242	56.58%	1.747
● FZBM	0.222	28.16%	4.748
● Code_NZ1	0.196	39.74%	1.976
● NA	0.193	41.32%	1.878
● Rf	0.168	22.37%	4.019
● Code_IA5	0.154	20.79%	3.834
● NTI	0.138	19.21%	3.571
● Code_N55	0.134	57.89%	1.307

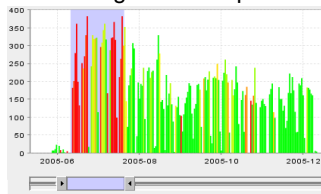
## Select split variable from a list

- Measure only needs to create a good ranking
- Very often, the expert-selected attribute is somewhat down the list

## Split quality in 2 dimensions



## Editing numeric splits



## Rule-based Subgroup Description

- Many rule induction approaches not suitable for subgroup description
    - incomplete
    - do not create synonyms
  - Hence we use complete search
    - but this entails redundancy problem
    - Filtering
      - ... by instance set overlap (e.g., Gebhardt)
      - ... by similarity of variables and values (e.g., Liu)
- too much based on statistics, and too little on expert knowledge

## Interactive Rule Sets

To bridge the gap, we let the user handle rule sets interactively

- Some conservative filtering may be done automatically
- Interactivity is supported by searching, sorting, grouping. . .
- Interactive filtering is provided by selecting rules and reducing weights of covered instances (similar to CN2-SD)
- Ongoing research concerned with further interactive filtering

## Conclusion

- We experienced a gap between theoretical ideas and practical application
- Interactivity: a key to close this gap
  - Simple modeling techniques
  - Experts apply them without assistance
  - Knowledge flows in both directions
- Data mining: rather insight provider than solution provider