Karlsruhe Institute of Technology

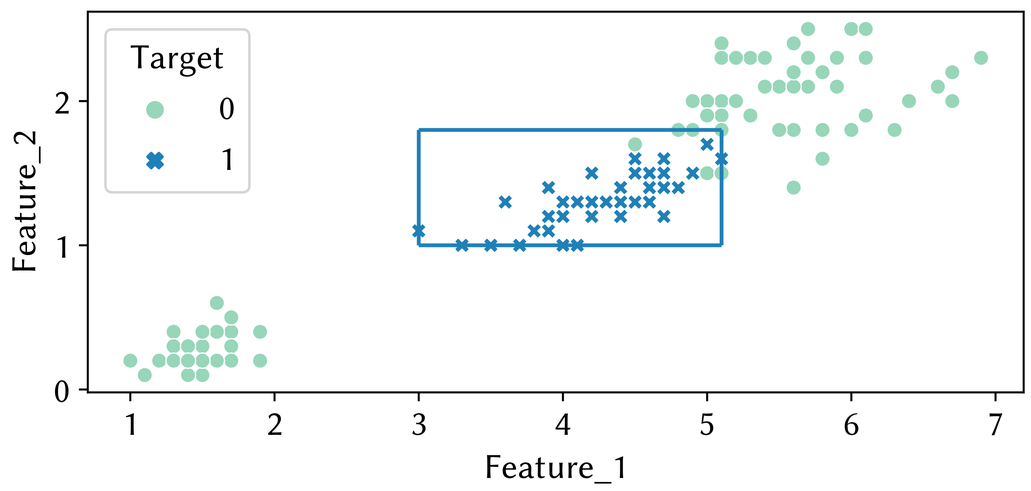
Subgroup Discovery with Small and Alternative Feature Sets

SIGMOD 2025 | Berlin

Jakob Bach (jakob.bach@kit.edu)

Scenario

- **Problem:** Find interesting, simple-to-describe region(s) in dataset



- **Our scope:** Binary classification with real-valued features
  - Tabular dataset $X \in \mathbb{R}^{m \times n}$ (data objects $\times$ features)
  - Prediction target $y \in \{0, 1\}^m$ ('interesting'/'positive' = 1)
  - Subgroup description: Hyperrectangle
  - Subgroup quality: Weighted Relative Accuracy (WRAcc)
- **Our focus**: Constraints for interpretable subgroup descriptions

Contributions

- Formalize subgroup discovery as an SMT optimization problem
- Formalize two constraint types (feature cardinality and alternatives)
- Analyze computational complexity and show $\mathcal{NP}$-hardness
- Comprehensive experiments

Formalization – Basic Problem

$$\max \quad Q_{\text{WRAcc}} = \frac{m_b}{m} \cdot \left(\frac{m_b^+}{m_b} - \frac{m^+}{m} \right) = \frac{m_b^+}{m} - \frac{m_b \cdot m^+}{m^2}$$

$$\text{s.t.:} \quad m_b := \sum_{i=1}^{m} b_i \quad \text{and} \quad m_b^+ := \sum_{\substack{i \in \{1,\ldots,m\} \\ y_i = 1}} b_i$$

$$\forall i \in \{1, \ldots, m\} \quad b_i \leftrightarrow \bigwedge_{j \in \{1,\ldots,n\}} \left((X_{ij} \geq lb_j) \wedge (X_{ij} \leq ub_j) \right)$$

$$\forall j \in \{1, \ldots, n\} \quad lb_j \leq ub_j$$

$$b \in \{0, 1\}^m$$

$$lb, ub \in \{\mathbb{R} \cup \{-\infty, +\infty\}\}^n$$

Paper

Code

Data

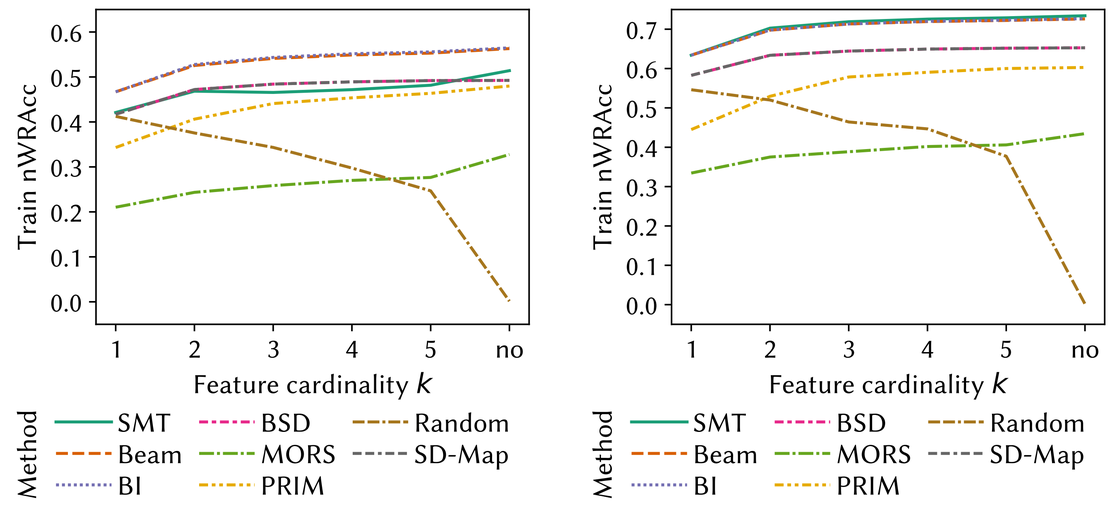
Constraints to Foster Interpretability

- **Feature-cardinality constraints:**
  - Limit number of features used in subgroup description to $k \in \mathbb{N}$
  - Feature used if its bounds exclude ≥ 1 data object from subgroup
- **Alternative subgroup descriptions:**
  - New optimization problem: Cover similar set of data objects as a given subgroup but use different feature set in description
  - Parameters: Number of alternatives a, dissimilarity threshold τ

Experimental Design

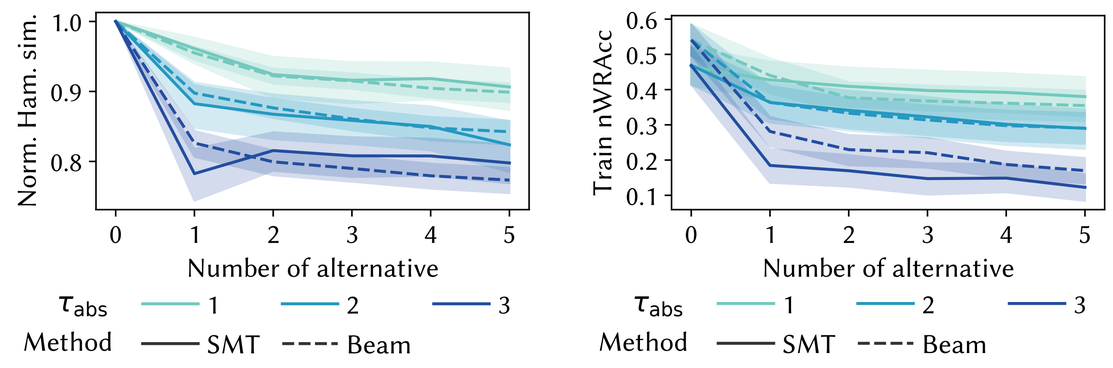
- 27 datasets from *PMLB* (with $m \in [106, 9822]$ and $n \in [20, 168]$)
- 8 subgroup-discovery methods: SMT-solver-based and 7 competitors
- Vary parameter values for constraints (k, a, τ) and solver's timeout

Experimental Results: Feature-Cardinality Constraints



- Left: all datasets; right: only datasets without solver timeouts
- Heuristics *Beam* and *BI* yield high subgroup quality (WRAcc) fast
- Using few features suffices to reach high subgroup quality

Experimental Results: Alternative Subgroup Descriptions



- Left: similarity to original subgroup; right: subgroup quality
- Similarity and quality of alternatives decreases over a and τ
- Strongest decrease from original subgroup to first alternative