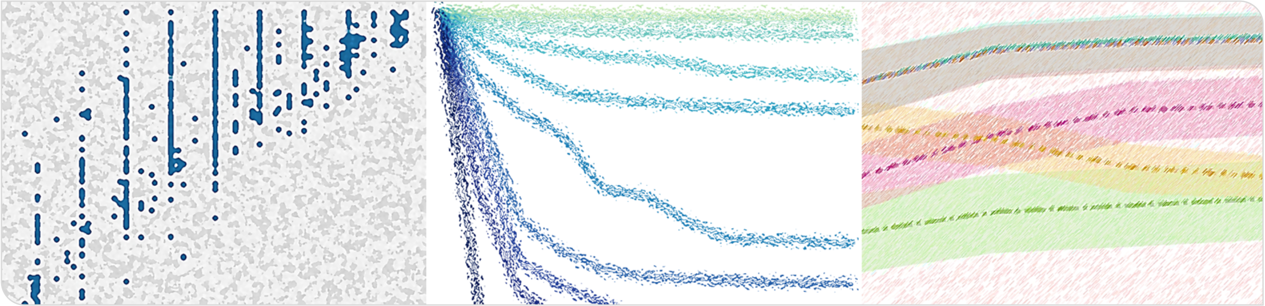
Karlsruhe Institute of Technology

Leveraging Constraints for User-Centric Feature Selection

PhD Defense

Jakob Bach | January 20, 2025



Background

Definition (Feature selection)

Background

Definition (Feature selection)

Given a dataset $X \in \mathbb{R}^{m \times n}$ with prediction target $y \in Y^m$ (e.g., $Y = \mathbb{R}$ or $Y = \{0, 1\}$),

Feat_1	Feat_2	...	Feat_n	Target
X_{11}	X_{12}	...	X_{1n}	y_1
X_{21}	X_{22}	...	X_{2n}	y_2
...	...	...	...	...
X_{m1}	X_{m2}	...	X_{mn}	y_m

Background

Karlsruhe Institute of Technology

Definition (Feature selection)

Given a dataset $X \in \mathbb{R}^{m \times n}$ with prediction target $y \in Y^m$ (e.g., $Y = \mathbb{R}$ or $Y = \{0, 1\}$), *feature selection* is the problem of making feature-selection decisions $s \in \{0, 1\}^n$

Feat_1	Feat_2	...	Feat_n	Target
X_{11}	X_{12}	...	X_{1n}	y_1
X_{21}	X_{22}	...	X_{2n}	y_2
...	...	...	...	...
X_{m1}	X_{m2}	...	X_{mn}	y_m

Background

Definition (Feature selection)

Given a dataset $X \in \mathbb{R}^{m \times n}$ with prediction target $y \in Y^m$ (e.g., $Y = \mathbb{R}$ or $Y = \{0, 1\}$), *feature selection* is the problem of making feature-selection decisions $s \in \{0, 1\}^n$ that maximize a given notion of feature-set quality $Q(s, X, y)$.

Feat_1	Feat_2	...	Feat_n	Target
X_{11}	X_{12}	...	X_{1n}	y_1
X_{21}	X_{22}	...	X_{2n}	y_2
...	...	...	...	...
X_{m1}	X_{m2}	...	X_{mn}	y_m

Background

Karlsruhe Institute of Technology

Definition (Feature selection)

Given a dataset $X \in \mathbb{R}^{m \times n}$ with prediction target $y \in Y^m$ (e.g., $Y = \mathbb{R}$ or $Y = \{0, 1\}$), *feature selection* is the problem of making feature-selection decisions $s \in \{0, 1\}^n$ that maximize a given notion of feature-set quality $Q(s, X, y)$.
Typically, select a fixed number of features $k \in \mathbb{N}$, i.e., $\sum_{j=1}^{n} s_j = k$.

Feat_1	Feat_2	...	Feat_n	Target
X_{11}	X_{12}	...	X_{1n}	y_1
X_{21}	X_{22}	...	X_{2n}	y_2
...	...	...	...	...
X_{m1}	X_{m2}	...	X_{mn}	y_m

Background

Definition (Feature selection)

Given a dataset $X \in \mathbb{R}^{m \times n}$ with prediction target $y \in Y^m$ (e.g., $Y = \mathbb{R}$ or $Y = \{0, 1\}$), *feature selection* is the problem of making feature-selection decisions $s \in \{0, 1\}^n$ that maximize a given notion of feature-set quality $Q(s, X, y)$.
Typically, select a fixed number of features $k \in \mathbb{N}$, i.e., $\sum_{j=1}^{n} s_j = k$.

- Reasons for feature selection [13, 34]:
  - Increase interpretability of predictions
  - Reduce computational requirements of machine learning (CPU, memory, storage, power consumption)
  - Improve prediction performance

Feat_1	Feat_2	...	Feat_n	Target
X_{11}	X_{12}	...	X_{1n}	y_1
X_{21}	X_{22}	...	X_{2n}	y_2
...	...	...	...	...
X_{m1}	X_{m2}	...	X_{mn}	y_m

Motivation and Our Approach

- Main limitations of most existing feature-selection methods:
  - Do not consider domain knowledge
  - Return only one feature set, no alternatives

Motivation and Our Approach

Karlsruhe Institute of Technology

- Main limitations of most existing feature-selection methods:
  - Do not consider domain knowledge
  - Return only one feature set, no alternatives
- Central idea of dissertation: Make feature selection more user-centric via constraints
  - Still optimize feature-set quality but restrict valid feature selections
  - Formulate as white-box optimization problem and use solver

Motivation and Our Approach

- Main limitations of most existing feature-selection methods:
  - Do not consider domain knowledge
  - Return only one feature set, no alternatives

- Central idea of dissertation: Make feature selection more user-centric via constraints
  - Still optimize feature-set quality but restrict valid feature selections
  - Formulate as white-box optimization problem and use solver

Example (A feature-selection constraint)

$(\neg s_1 \land \neg s_2 \land \neg s_3) \lor (s_1 \land s_2 \land s_3) \leftrightarrow$ "Select none or all of Features 1, 2, and 3."

Motivation and Our Approach

- Main limitations of most existing feature-selection methods:
  - Do not consider domain knowledge
  - Return only one feature set, no alternatives

- Central idea of dissertation: Make feature selection more user-centric via constraints
  - Still optimize feature-set quality but restrict valid feature selections
  - Formulate as white-box optimization problem and use solver

Example (A feature-selection constraint)

$(\neg s_1 \land \neg s_2 \land \neg s_3) \lor (s_1 \land s_2 \land s_3) \leftrightarrow$ "Select none or all of Features 1, 2, and 3."

- Benefits of our approach:
  - Declarative
  - Allows combining constraints
  - Orthogonal to choice of feature-selection method

Our Contributions

- (C1) Evaluating the impact of constraints [7]
  - Formalize constrained feature selection
  - Conduct domain-independent study on impact of constraints
- (C2) Using constraints to formulate scientific hypotheses [7]
  - Conduct domain-specific study on impact of constraints

Our Contributions

- (C1) Evaluating the impact of constraints [7]
  - Formalize constrained feature selection
  - Conduct domain-independent study on impact of constraints
- (C2) Using constraints to formulate scientific hypotheses [7]
  - Conduct domain-specific study on impact of constraints
- (C3) Using constraints for alternative feature sets [3, 6]
  - Formalize alternative feature selection
  - Discuss integrating constraints
  - Analyze time complexity
  - Propose heuristic search methods
  - Conduct experimental study

Our Contributions

Karlsruhe Institute of Technology

- (C1) Evaluating the impact of constraints [7]
  - Formalize constrained feature selection
  - Conduct domain-independent study on impact of constraints
- (C2) Using constraints to formulate scientific hypotheses [7]
  - Conduct domain-specific study on impact of constraints
- (C3) Using constraints for alternative feature sets [3, 6]
  - Formalize alternative feature selection
  - Discuss integrating constraints
  - Analyze time complexity
  - Propose heuristic search methods
  - Conduct experimental study
- (C4) Using constraints for feature selection in subgroup discovery [4, 5]
  - Formalize subgroup discovery as SMT optimization problem
  - Formalize two constraint types
  - Discuss integrating constraints
  - Analyze time complexity
  - Conduct experimental study

Our Contributions

- (C1) Evaluating the impact of constraints [7]
  - Formalize constrained feature selection
  - Conduct domain-independent study on impact of constraints
- (C2) Using constraints to formulate scientific hypotheses [7]
  - Conduct domain-specific study on impact of constraints
- (C3) Using constraints for alternative feature sets [3, 6]
  - Formalize alternative feature selection
  - Discuss integrating constraints
  - Analyze time complexity
  - Propose heuristic search methods
  - Conduct experimental study
- (C4) Using constraints for feature selection in subgroup discovery [4, 5]
  - Formalize subgroup discovery as SMT optimization problem
  - Formalize two constraint types
  - Discuss integrating constraints
  - Analyze time complexity
  - Conduct experimental study
- Reproducibility:
  - All experimental data available on RADAR4KIT
  - Three GitHub repositories [a, b, c]
  - Three Python packages: `alfese`, `cffs`, `csd`

(C1) Evaluating the Impacts of Constraints – Approach

- Systematic study on impact of constraints on feature-selection results
  - Metrics for constraints, e.g., fraction of valid feature sets
  - Metrics for results, e.g., feature-set quality

(C1) Evaluating the Impacts of Constraints – Approach

- Systematic study on impact of constraints on feature-selection results
  - Metrics for constraints, e.g., fraction of valid feature sets
  - Metrics for results, e.g., feature-set quality

- Experimental design:
  - 35 regression datasets from *OpenML* [53]
  - Linear objective: $Q(s, X, y) = \sum_{j=1}^{n} q(X_{\cdot j}, y) \cdot s_j$ (using mutual information [29] as $q(\cdot)$)
  - Generate random constraints for ten constraint types with 1000 repetitions
  - *Z3* [10, 16] (an SMT solver) as optimizer

(C1) Evaluating the Impacts of Constraints – Approach

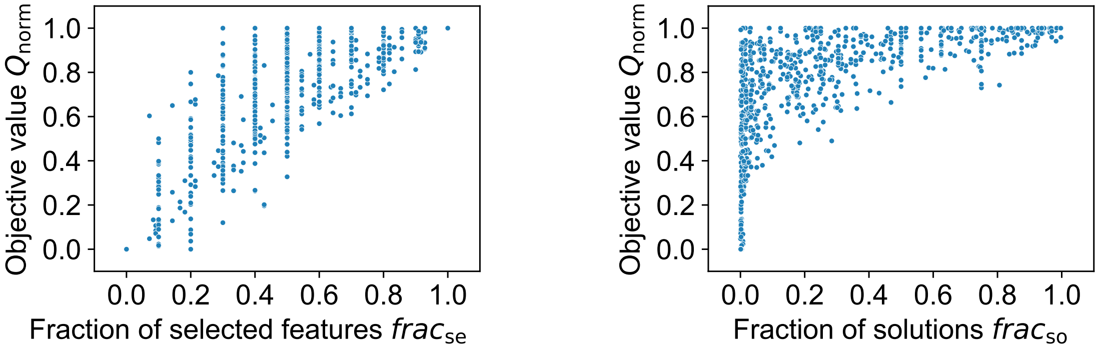
- Systematic study on impact of constraints on feature-selection results
  - Metrics for constraints, e.g., fraction of valid feature sets
  - Metrics for results, e.g., feature-set quality

- Experimental design:
  - 35 regression datasets from *OpenML* [53]
  - Linear objective: $Q(s, X, y) = \sum_{j=1}^{n} q(X_{\cdot j}, y) \cdot s_j$ (using mutual information [29] as $q(\cdot)$)
  - Generate random constraints for ten constraint types with 1000 repetitions
  - *Z3* [10, 16] (an SMT solver) as optimizer

Examples (Constraint types)

- Single-XOR$(s_{j_1}, s_{j_2}) = s_{j_1} \oplus s_{j_2} = (s_{j_1} \land \neg s_{j_2}) \lor (\neg s_{j_1} \land s_{j_2})$
- Group-NAND$(\{s_{j_1}, \ldots, s_{j_{n'}}\}) = \neg(s_{j_1} \land s_{j_2} \land \cdots \land s_{j_{n'}}) = \sum_{l=1}^{n'} s_{j_l} \leq n' - 1$

(C1) Evaluating the Impacts of Constraints – Results



$$frac_{\text{se}} = \frac{\sum_{j=1}^{n} s_j}{n} \quad \Bigg| \quad Q_{\text{norm}} = \frac{\sum_{j=1}^{n} s_j \cdot q(X_{\cdot j}, y)}{\sum_{j=1}^{n} q(X_{\cdot j}, y)} \quad \Bigg| \quad frac_{\text{so}} = \frac{\sum_{s \in \{0,1\}^n} \min\limits_{c \in C} c(s)}{2^n}$$

(C3) Alternative Feature Selection – Approach

- Idea: Find feature sets optimizing feature-set quality while differing from each other
  - Domain-independent constraint type
  - Orthogonal to choice of feature-selection method
  - Number of alternatives $a \in \mathbb{N}_0$ and feature-set dissimilarity threshold $\tau \in [0, 1]$ as user parameters
  - Sequential or simultaneous search

(C3) Alternative Feature Selection – Approach

- Idea: Find feature sets optimizing feature-set quality while differing from each other
  - Domain-independent constraint type
  - Orthogonal to choice of feature-selection method
  - Number of alternatives $a \in \mathbb{N}_0$ and feature-set dissimilarity threshold $\tau \in [0, 1]$ as user parameters
  - Sequential or simultaneous search

Sequential-search problem

$$\max_{s} \quad Q(s, X, y)$$

$$\text{subject to:} \quad \forall F' \in \mathbb{F}: \; d(F_s, F') \geq \tau$$

Simultaneous-search problem

$$\max_{s^{(0)},\ldots,s^{(a)}} \quad \underset{l \in \{0,\ldots,a\}}{\operatorname{agg}} \quad Q(s^{(l)}, X, y)$$

$$\text{subject to:} \quad \forall l_1, l_2 \in \{0, \ldots, a\},\; l_1 \neq l_2: \; d(F_{s^{(l_1)}}, F_{s^{(l_2)}}) \geq \tau$$

(C3) Alternative Feature Selection – Approach

- Idea: Find feature sets optimizing feature-set quality while differing from each other
  - Domain-independent constraint type
  - Orthogonal to choice of feature-selection method
  - Number of alternatives $a \in \mathbb{N}_0$ and feature-set dissimilarity threshold $\tau \in [0, 1]$ as user parameters
  - Sequential or simultaneous search

Sequential-search problem

$$\max_{s} \quad Q(s, X, y)$$
$$\text{subject to:} \quad \forall F' \in \mathbb{F}: \; d(F_s, F') \geq \tau$$

Simultaneous-search problem

$$\max_{s^{(0)},\dots,s^{(a)}} \quad \underset{l \in \{0,\dots,a\}}{\text{agg}} \quad Q(s^{(l)}, X, y)$$
$$\text{subject to:} \quad \forall l_1, l_2 \in \{0, \dots, a\}, \; l_1 \neq l_2: \; d(F_{s^{(l_1)}}, F_{s^{(l_2)}}) \geq \tau$$

- Chosen dissimilarity measure: $d_{\text{Dice}}(F', F'') = 1 - \frac{2 \cdot |F' \cap F''|}{|F'| + |F''|} = 1 - \frac{2 \cdot \sum_{j=1}^{n} s'_j \cdot s''_j}{\sum_{j=1}^{n} s'_j + \sum_{j=1}^{n} s''_j}$

(C3) Alternative Feature Selection – Complexity

- Alternative feature selection is $\mathcal{NP}$-complete in scenario with:
  - Simultaneous search with minimum as aggregation operator agg$(\cdot)$
  - Linear notion of feature-set quality: $Q(s, X, y) = \sum_{j=1}^{n} q_j \cdot s_j$
  - No feature-set overlap ($\tau = 1$)
  - Complete partitioning (each feature part of exactly one feature set)

(C3) Alternative Feature Selection – Complexity

Karlsruhe Institute of Technology

- Alternative feature selection is $\mathcal{NP}$-complete in scenario with:
  - Simultaneous search with minimum as aggregation operator agg$(\cdot)$
  - Linear notion of feature-set quality: $Q(s, X, y) = \sum_{j=1}^{n} q_j \cdot s_j$
  - No feature-set overlap ($\tau = 1$)
  - Complete partitioning (each feature part of exactly one feature set)
- Proof: Known as MULTI-WAY NUMBER PARTITIONING [28] or MULTIPROCESSOR SCHEDULING [22] □

(C3) Alternative Feature Selection – Complexity

- Alternative feature selection is $\mathcal{NP}$-complete in scenario with:
  - Simultaneous search with minimum as aggregation operator agg($\cdot$)
  - Linear notion of feature-set quality: $Q(s, X, y) = \sum_{j=1}^{n} q_j \cdot s_j$
  - No feature-set overlap ($\tau = 1$)
  - Complete partitioning (each feature part of exactly one feature set)
- Proof: Known as MULTI-WAY NUMBER PARTITIONING [28] or MULTIPROCESSOR SCHEDULING [22] □

- Further complexity-related contributions and results:
  - Prove $\mathcal{NP}$-hardness for (1) incomplete partitioning and (2) feature-set overlap

(C3) Alternative Feature Selection – Complexity

- Alternative feature selection is $\mathcal{NP}$-complete in scenario with:
  - Simultaneous search with minimum as aggregation operator agg($\cdot$)
  - Linear notion of feature-set quality: $Q(s, X, y) = \sum_{j=1}^{n} q_j \cdot s_j$
  - No feature-set overlap ($\tau = 1$)
  - Complete partitioning (each feature part of exactly one feature set)
- Proof: Known as MULTI-WAY NUMBER PARTITIONING [28] or MULTIPROCESSOR SCHEDULING [22] □

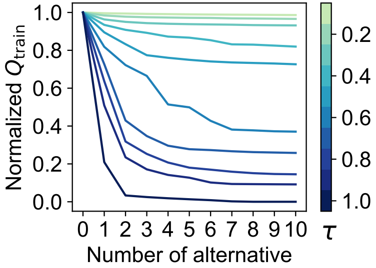
- Further complexity-related contributions and results:
  - Prove $\mathcal{NP}$-hardness for (1) incomplete partitioning and (2) feature-set overlap
  - Prove polynomial runtime for sum-aggregation and sequential search with remaining conditions from above

(C3) Alternative Feature Selection – Complexity

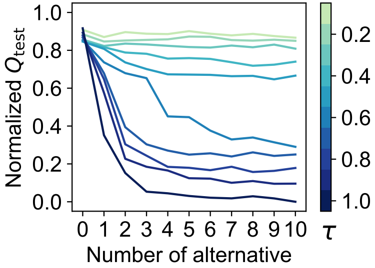
- Alternative feature selection is $\mathcal{NP}$-complete in scenario with:
  - Simultaneous search with minimum as aggregation operator $\text{agg}(\cdot)$
  - Linear notion of feature-set quality: $Q(s, X, y) = \sum_{j=1}^{n} q_j \cdot s_j$
  - No feature-set overlap ($\tau = 1$)
  - Complete partitioning (each feature part of exactly one feature set)
- Proof: Known as MULTI-WAY NUMBER PARTITIONING [28] or MULTIPROCESSOR SCHEDULING [22] □

- Further complexity-related contributions and results:
  - Prove $\mathcal{NP}$-hardness for (1) incomplete partitioning and (2) feature-set overlap
  - Prove polynomial runtime for sum-aggregation and sequential search with remaining conditions from above
  - Develop two heuristic search methods with approximation guarantee

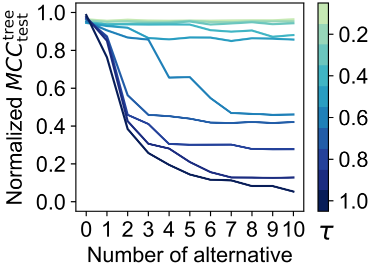
(C3) Alternative Feature Selection – Empirical Results



(a) Training-set objective value.



(b) Test-set objective value.

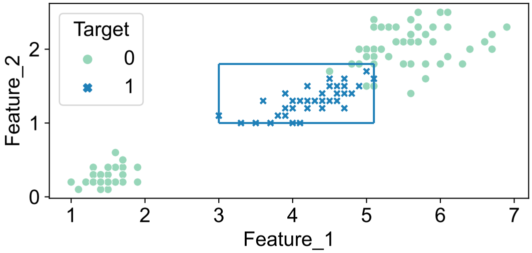


(c) Test-set prediction performance.

Mean of feature-set quality, over the number of alternatives and dissimilarity threshold τ, by evaluation metric.

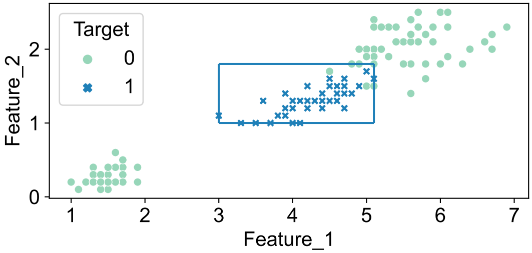
(C4) Constrained Subgroup Discovery – Approach

- Subgroup discovery: “Identifying descriptions of subsets of a dataset that show an interesting behavior” [1]



(C4) Constrained Subgroup Discovery – Approach

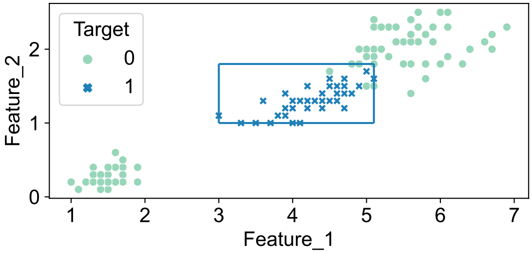
- Subgroup discovery: “Identifying descriptions of subsets of a dataset that show an interesting behavior” [1]
- Scope: Binary classification with real-valued features
  - Tabular dataset $X \in \mathbb{R}^{m \times n}$ (data objects $\times$ features)
  - Prediction target $y \in \{0, 1\}^m$ (‘interesting’/‘positive’ = 1)



(C4) Constrained Subgroup Discovery – Approach

- Subgroup discovery: "Identifying descriptions of subsets of a dataset that show an interesting behavior" [1]

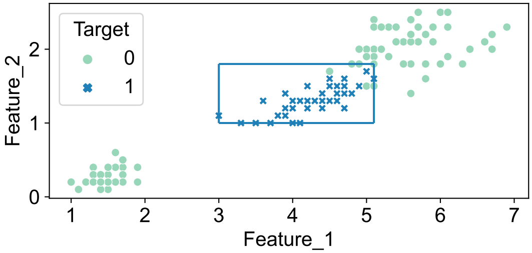
- Scope: Binary classification with real-valued features
  - Tabular dataset $X \in \mathbb{R}^{m \times n}$ (data objects $\times$ features)
  - Prediction target $y \in \{0, 1\}^m$ ('interesting'/'positive' = 1)
  - Subgroup quality: Weighted Relative Accuracy
    - $\text{WRAcc} = \frac{m_b}{m} \cdot \left(\frac{m_b^+}{m_b} - \frac{m^+}{m} \right)$ [31]
    - $+ \leftrightarrow$ positive data object, $b \leftrightarrow$ in subgroup (box)



(C4) Constrained Subgroup Discovery – Approach

- Subgroup discovery: "Identifying descriptions of subsets of a dataset that show an interesting behavior" [1]

- 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 quality: Weighted Relative Accuracy
    - $\text{WRAcc} = \frac{m_b}{m} \cdot \left(\frac{m_b^+}{m_b} - \frac{m^+}{m} \right)$ [31]
    - $+ \leftrightarrow$ positive data object, $b \leftrightarrow$ in subgroup (box)



- Our goal: Improve interpretability with constraints
  - Limit number of used features
  - Find alternative subgroup descriptions

(C4) Subgroup Discovery – Formalization: Basic Problem

$$lb, ub \in \{\mathbb{R} \cup \{-\infty, +\infty\}\}^n \qquad \text{(Variables: lower/upper bounds of subgroup)}$$

(C4) Subgroup Discovery – Formalization: Basic Problem

$$\forall j \in \{1, \ldots, n\} \quad lb_j \leq ub_j \qquad \text{(Constraint: relationship between bounds)}$$

$$lb, ub \in \{\mathbb{R} \cup \{-\infty, +\infty\}\}^n \qquad \text{(Variables: lower/upper bounds of subgroup)}$$

(C4) Subgroup Discovery – Formalization: Basic Problem

$$\forall i \in \{1, \dots, m\} \quad b_i \leftrightarrow \bigwedge_{j \in \{1,\dots,n\}} \left((X_{ij} \geq lb_j) \land (X_{ij} \leq ub_j) \right) \quad \text{(i-th data object in subgroup?)}$$

$$\forall j \in \{1, \dots, n\} \quad lb_j \leq ub_j \quad \text{(Constraint: relationship between bounds)}$$

$$b \in \{0, 1\}^m \quad \text{(Auxiliary variables: subgroup membership)}$$

$$lb, ub \in \{\mathbb{R} \cup \{-\infty, +\infty\}\}^n \quad \text{(Variables: lower/upper bounds of subgroup)}$$

(C4) Subgroup Discovery – Formalization: Basic Problem

s.t.: $m_b := \sum_{i=1}^{m} b_i$ and $m_b^+ := \sum_{\substack{i \in \{1,\dots,m\} \\ y_i = 1}} b_i$ (Num of data objects in subgroup)

$\forall i \in \{1, \dots, m\}$ $b_i \leftrightarrow \bigwedge_{j \in \{1,\dots,n\}} ((X_{ij} \geq lb_j) \land (X_{ij} \leq ub_j))$ (i-th data object in subgroup?)

$\forall j \in \{1, \dots, n\}$ $lb_j \leq ub_j$ (Constraint: relationship between bounds)

$b \in \{0, 1\}^m$ (Auxiliary variables: subgroup membership)

$lb, ub \in \{\mathbb{R} \cup \{-\infty, +\infty\}\}^n$ (Variables: lower/upper bounds of subgroup)

(C4) Subgroup Discovery – Formalization: Basic Problem

Karlsruhe Institute of Technology

max $\quad Q_{\text{WRAcc}} = \frac{m_b^+}{m} - \frac{m_b \cdot m^+}{m^2}$ (Objective: subgroup quality)

s.t.: $\quad m_b := \sum_{i=1}^{m} b_i$ and $m_b^+ := \sum_{\substack{i \in \{1,\dots,m\} \\ y_i = 1}} b_i$ (Num of data objects in subgroup)

$\forall i \in \{1, \dots, m\} \quad b_i \leftrightarrow \bigwedge_{j \in \{1,\dots,n\}} \left((X_{ij} \geq lb_j) \land (X_{ij} \leq ub_j) \right)$ (i-th data object in subgroup?)

$\forall j \in \{1, \dots, n\} \quad lb_j \leq ub_j$ (Constraint: relationship between bounds)

$b \in \{0, 1\}^m$ (Auxiliary variables: subgroup membership)

$lb, ub \in \{\mathbb{R} \cup \{-\infty, +\infty\}\}^n$ (Variables: lower/upper bounds of subgroup)

(C4) Subgroup Discovery – Formalization: Cardinality

- Concept: Limit number of features used (= selected) in subgroup description to $k \in \mathbb{N}$
- Formally: Limit number of features whose bounds exclude at least one data object from subgroup

(C4) Subgroup Discovery – Formalization: Cardinality

- Concept: Limit number of features used (= selected) in subgroup description to $k \in \mathbb{N}$
- Formally: Limit number of features whose bounds exclude at least one data object from subgroup

$$s, s^{\text{lb}}, s^{\text{ub}} \in \{0, 1\}^n$$

(C4) Subgroup Discovery – Formalization: Cardinality

- Concept: Limit number of features used (= selected) in subgroup description to $k \in \mathbb{N}$
- Formally: Limit number of features whose bounds exclude at least one data object from subgroup

$$\forall j \in \{1, \ldots, n\}: \quad s_j^{\text{lb}} \leftrightarrow \left(lb_j > \min_{i \in \{1, \ldots, m\}} X_{ij} \right)$$

$$\forall j \in \{1, \ldots, n\}: \quad s_j^{\text{ub}} \leftrightarrow \left(ub_j < \max_{i \in \{1, \ldots, m\}} X_{ij} \right)$$

$$s, s^{\text{lb}}, s^{\text{ub}} \in \{0, 1\}^n$$

(C4) Subgroup Discovery – Formalization: Cardinality

- Concept: Limit number of features used (= selected) in subgroup description to $k \in \mathbb{N}$
- Formally: Limit number of features whose bounds exclude at least one data object from subgroup

$$\forall j \in \{1, \dots, n\}: \quad s_j^{\text{lb}} \leftrightarrow \left(lb_j > \min_{i \in \{1, \dots, m\}} X_{ij} \right)$$

$$\forall j \in \{1, \dots, n\}: \quad s_j^{\text{ub}} \leftrightarrow \left(ub_j < \max_{i \in \{1, \dots, m\}} X_{ij} \right)$$

$$\forall j \in \{1, \dots, n\}: \quad s_j \leftrightarrow \left(s_j^{\text{lb}} \lor s_j^{\text{ub}} \right)$$

$$s, s^{\text{lb}}, s^{\text{ub}} \in \{0, 1\}^n$$

(C4) Subgroup Discovery – Formalization: Cardinality

- Concept: Limit number of features used (= selected) in subgroup description to $k \in \mathbb{N}$
- Formally: Limit number of features whose bounds exclude at least one data object from subgroup

$$\forall j \in \{1, \ldots, n\}: \quad s_j^{\text{lb}} \leftrightarrow \left(lb_j > \min_{i \in \{1, \ldots, m\}} X_{ij} \right)$$

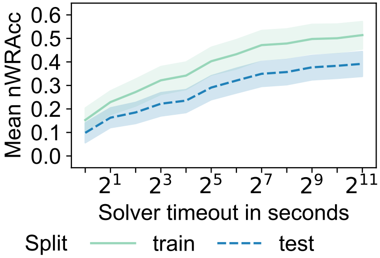
$$\forall j \in \{1, \ldots, n\}: \quad s_j^{\text{ub}} \leftrightarrow \left(ub_j < \max_{i \in \{1, \ldots, m\}} X_{ij} \right)$$

$$\forall j \in \{1, \ldots, n\}: \quad s_j \leftrightarrow \left(s_j^{\text{lb}} \lor s_j^{\text{ub}} \right)$$

$$\sum_{j=1}^{n} s_j \leq k$$

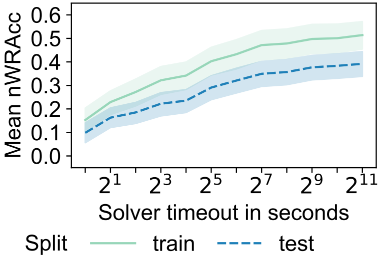
$$s, s^{\text{lb}}, s^{\text{ub}} \in \{0, 1\}^n$$

(C4) Constrained Subgroup Discovery – Empirical Results

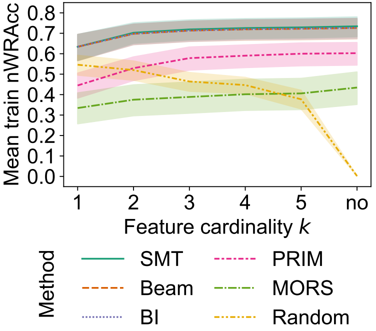


Mean subgroup quality over solver timeouts for *SMT* search.

(C4) Constrained Subgroup Discovery – Empirical Results

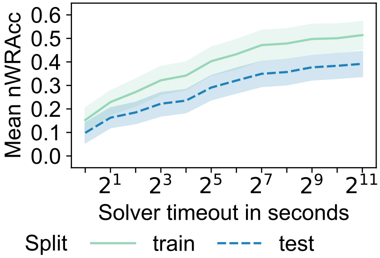


Mean subgroup quality over solver timeouts for *SMT* search.

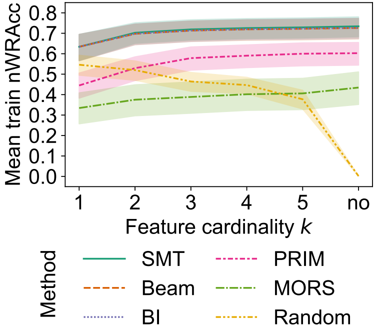


Mean training-set subgroup quality over feature-cardinality threshold k.

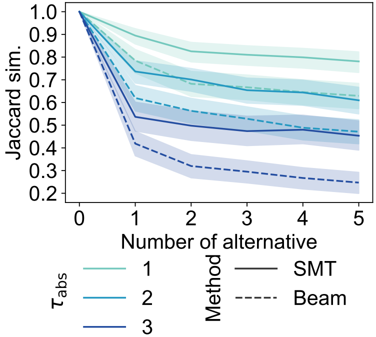
(C4) Constrained Subgroup Discovery – Empirical Results



Mean subgroup quality over solver timeouts for *SMT* search.



Mean training-set subgroup quality over feature-cardinality threshold *k*.



Mean similarity of alternative subgroup descriptions.

Conclusions

- Research gaps – Most existing methods for feature selection:
  - Do not consider domain knowledge
  - Return only one solution, no alternatives

Conclusions

- Research gaps – Most existing methods for feature selection:
  - Do not consider domain knowledge
  - Return only one solution, no alternatives

- Approach: Address both research gaps with constraints on selected feature sets

Conclusions

- Research gaps – Most existing methods for feature selection:
  - Do not consider domain knowledge
  - Return only one solution, no alternatives
- Approach: Address both research gaps with constraints on selected feature sets
- Core contributions:
  - (C1) Evaluating the impact of constraints: Bach et al. (2022) [7]
  - (C2) Using constraints to formulate scientific hypotheses: Bach et al. (2022) [7]
  - (C3) Using constraints for alternative feature sets: Bach (2024) [3], Bach and Böhm (2024) [6]
  - (C4) Using constraints for feature selection in subgroup discovery: Bach (2025) [4], Bach (2024) [5]
- All code and data available online

Appendix

Details of Our Underlying Publications

- Constrained feature selection (C1 and C2):
  - Jakob Bach et al. "An Empirical Evaluation of Constrained Feature Selection". In: *SN Comput. Sci.* 3.6 (2022). DOI: 10.1007/s42979-022-01338-z
- Alternative feature selection (C3):
  - Jakob Bach. *Finding Optimal Diverse Feature Sets with Alternative Feature Selection*. arXiv:2307.11607v2 [cs.LG]. 2024. DOI: 10.48550/arXiv.2307.11607
  - Jakob Bach and Klemens Böhm. "Alternative feature selection with user control". In: *Int. J. Data Sci. Anal.* (2024). DOI: 10.1007/s41060-024-00527-8
- Constrained subgroup discovery (C4):
  - Jakob Bach. *Using Constraints to Discover Sparse and Alternative Subgroup Descriptions*. arXiv:2406.01411v1 [cs.LG]. 2024. DOI: 10.48550/arXiv.2406.01411
  - Jakob Bach. *Subgroup Discovery with Small and Alternative Feature Sets*. Conditionally accepted at SIGMOD 2025.

Related Work

- Integrating domain knowledge and constraints:
  - Feature selection: Typically only combination of one constraint type (like cost [45, 47], cardinality [27, 54], or group [25, 55]) and feature-selection method; exceptions (wrapper methods with black-box constraints): [23, 42]
  - Subgroup discovery: White-box formulations of different problem definitions [18, 35] and integration of constraints into algorithmic search methods [2, 40]
  - Other fields: E.g., AutoML [43], clustering [15], pattern mining [51], XAI [17]; outside ML: software engineering [21]

- Finding alternative solutions:
  - Feature selection: Approaches that offer less user control over alternatives, e.g., ensemble feature selection [24, 50] or statistically equivalent feature sets [12, 30]
  - Subgroup discovery: Subgroup-set selection [36, 48]; different problem definitions of alternative descriptions, e.g., description-based subgroup selection [33] or equivalent subgroup descriptions of minimal length [11]
  - Other fields: E.g., clustering [8], number partitioning [32], subspace search [19], XAI [41]

Future Work

- Different areas/directions: ML methodology, applied ML, and complexity theory

- ML methodology:
  - Integrating constraints into more methods
  - Soft constraints
  - Feature engineering
- Applied ML:
  - Case studies with qualitative interpretation of results
  - User-friendly systems
- Complexity theory:
  - Approximation complexity
  - Parameterized complexity

(C2) Scientific Hypotheses as Constraints – Approach

Karlsruhe Institute of Technology

- Case study on impact of constraints in a specific use case
  - Scenario [52] from materials science: Predict density of dislocation reactions in a specimen under load
  - Constraints express preferences regarding feature sets and hypotheses from domain
  - Idea: Hypotheses inconsistent to data should lower prediction quality significantly

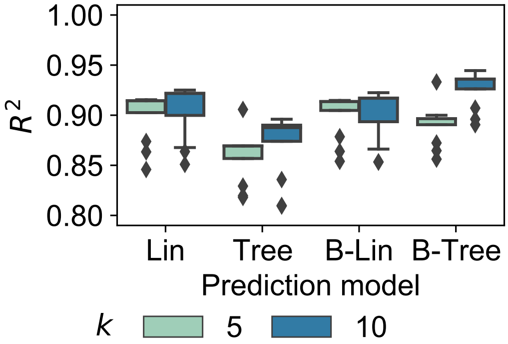
- Experimental design:
  - One dataset with 14,903 data objects, 135 features, and continuous target
  - Linear objective using Pearson correlation as $q(\cdot)$, four prediction models
  - Constraint types: Three domain-independent (preferences) and twelve domain-specific (hypotheses)

Example (Constraint type)

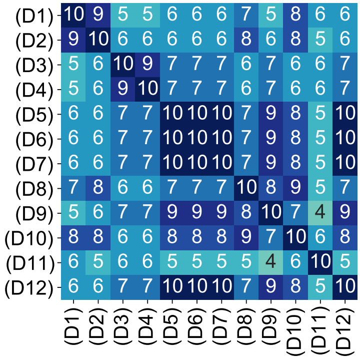
$$\text{Aggregate-or-original}(\{s_1, \ldots, s_n\}) = \bigwedge_{p \in P} \left(\left(\bigvee_{a \in A} s_{(p,a)} \right) + \left(\bigvee_{l \in \{1,\ldots,12\}} s_{(p,l)} \right) \leq 1 \right)$$

$\leftrightarrow$ "For each physical quantity p, do not select aggregate features a and original features l at the same time."

(C2) Scientific Hypotheses as Constraints – Results



Distribution of prediction quality over constraint types (which express scientific hypotheses).



Overlap size of feature sets between constraint types (for selecting $k = 10$ features).

(C3) Alternative Feature Selection – Experimental Design

- 30 binary-classification datasets (with 106–9822 data objects and 15–168 features) from *PMLB* [44, 49]
- Five feature-set quality measures as objective functions
- Different search configurations for alternatives:
  - Number of alternatives *a*
  - Dissimilarity threshold τ
  - Search methods (three solver-based, two heuristics)
- Evaluation metrics:
  - Objective value
  - Prediction performance (MCC [39])
  - Runtime
  - Optimization status
- *SCIP* [9] (a MIP solver) via Google OR-Tools [46] as optimizer for solver-based search

(C4) Subgroup Discovery – Formalization: Alternatives

- Concept: Cover similar set of data objects as given subgroup with different set of selected features
  - Repeat sequentially to get $a \in \mathbb{N}$ alternatives for dissimilarity threshold $\tau \in \mathbb{R}_{\geq 0}$

- Chosen optimization objective: Maximize normalized Hamming similarity (= prediction accuracy)

$$\text{sim}_{\text{nHamm}}(b^{(a)}, b^{(0)}) = \frac{1}{m} \cdot \sum_{i=1}^{m} \left(b_i^{(a)} \leftrightarrow b_i^{(0)} \right) = \frac{1}{m} \cdot \Bigg(\sum_{\substack{i \in \{1,\dots,m\} \\ b_i^{(0)}=1}} b_i^{(a)} + \sum_{\substack{i \in \{1,\dots,m\} \\ b_i^{(0)}=0}} \neg b_i^{(a)} \Bigg)$$

- Chosen dissimilarity constraints: From each existing feature set, deselect at least $\tau_{\text{abs}} \in \mathbb{N}$ (but $\leq k$) features

$$\forall l \in \{0, \dots, a-1\}: \; \text{dis}_{\text{des}}(s^{(a)}, s^{(l)}) = \sum_{\substack{j \in \{1,\dots,n\} \\ s_j^{(l)}=1}} \neg s_j^{(a)} \geq \min\left(\tau_{\text{abs}}, \; k^{(l)}\right)$$

(C4) Subgroup Discovery – Complexity

- Subgroup discovery with a feature-cardinality constraint is $\mathcal{NP}$-complete
- Proof: Reduction from SET COVERING [26]
  - Set-covering problem: Given set of elements $E = \{e_1, \dots, e_m\}$, set of sets $\mathbb{S} = \{S_1, \dots, S_n\}$ with $E = \bigcup_{S \in \mathbb{S}} S$, and a cardinality $k \in \mathbb{N}$, does subset $\mathbb{C} \subseteq \mathbb{S}$ with $|\mathbb{C}| \leq k$ and $E = \bigcup_{S \in \mathbb{C}} S$ exist?
  - Perfect-subgroup discovery: Find subgroup containing all positive data objects ($y_i = 1$) and zero negatives ($y_i = 0$)
  - Problem transformation:
    - Dataset $X \in \{0, 1\}^{(m+1) \times n}$ with $X_{ij} := (e_i \in S_j)$
    - Data Object $m + 1$ represents an element not contained in any set, i.e., $X_{(m+1)j} = 0$
    - Prediction target $y \in \{0, 1\}^{m+1}$ with $y_{m+1} = 1$ and $y_i = 0$ otherwise
  - Perfect subgroup only contains Data Object $m + 1$ and uses $lb_j = ub_j = 0$ conditions for selected features
  - Other data objects have value 1 for at least one selected feature $\rightarrow$ each element is in a selected set
  - I.e., algorithm for perfect-subgroup discovery also solves SET COVERING
  - Finally, optimizing subgroup quality typically at least as hard as finding perfect subgroup

(C4) Subgroup Discovery – Experimental Design

Karlsruhe Institute of Technology

- 27 binary-classification datasets (with 106–9822 data objects and 20–168 features) from *PMLB* [44, 49]
- Six subgroup-discovery methods:
  - Solver-based (novel): *SMT* (using *Z3* [10, 16] as optimizer)
  - Heuristics (related work): *Beam*, *BI* [37], *PRIM* [20]
  - Baselines (novel): *MORS* (Minimal Optimal Recall Subgroup), *Random*
- Four experimental scenarios: Unconstrained, two constraint types, and solver timeouts
  - Solver timeouts: {1 s, 2 s, 4 s, . . . , 2048 s}
  - Feature-cardinality constraints: $k \in \{1, 2, 3, 4, 5\}$
  - Alternative subgroup descriptions: $k = 3$, $a = 5$, and $\tau_{\text{abs}} \in \{1, 2, 3\}$
- Evaluation metrics:
  - Subgroup quality (nWRAcc [31, 38])
  - Runtime
  - For alternatives: Similarity [14] (Normalized Hamming and Jaccard)

References I

[1] Martin Atzmueller. “Subgroup discovery”. In: *WIREs Data Min. Knowl. Disc.* 5.1 (2015), pp. 35–49. DOI: 10.1002/widm.1144.

[2] Martin Atzmueller, Frank Puppe, and Hans-Peter Buscher. “Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery”. In: *Proc. IJCAI*. Edinburgh, United Kingdom, 2005, pp. 647–652. URL: https://www.ijcai.org/Proceedings/05/Papers/1217.pdf.

[3] Jakob Bach. *Finding Optimal Diverse Feature Sets with Alternative Feature Selection*. arXiv:2307.11607v2 [cs.LG]. 2024. DOI: 10.48550/arXiv.2307.11607.

[4] Jakob Bach. *Subgroup Discovery with Small and Alternative Feature Sets*. Conditionally accepted at SIGMOD 2025.

[5] Jakob Bach. *Using Constraints to Discover Sparse and Alternative Subgroup Descriptions*. arXiv:2406.01411v1 [cs.LG]. 2024. DOI: 10.48550/arXiv.2406.01411.

[6] Jakob Bach and Klemens Böhm. “Alternative feature selection with user control”. In: *Int. J. Data Sci. Anal.* (2024). DOI: 10.1007/s41060-024-00527-8.

References II

[7] Jakob Bach et al. “An Empirical Evaluation of Constrained Feature Selection”. In: *SN Comput. Sci.* 3.6 (2022). DOI: 10.1007/s42979-022-01338-z.

[8] James Bailey. “Alternative Clustering Analysis: A Review”. In: *Data Clustering: Algorithms and Applications*. 1st ed. CRC Press, 2014. Chap. 21, pp. 535–550. DOI: 10.1201/9781315373515.

[9] Ksenia Bestuzheva et al. *The SCIP Optimization Suite 8.0*. Tech. rep. Zuse Institute Berlin, 2021. URL: http://nbn-resolving.de/urn:nbn:de:0297-zib-85309.

[10] Nikolaj Bjørner, Anh-Dung Phan, and Lars Fleckenstein. “νZ - An Optimizing SMT Solver”. In: *Proc. TACAS*. London, United Kingdom, 2015, pp. 194–199. DOI: 10.1007/978-3-662-46681-0_14.

[11] Mario Boley and Henrik Grosskreutz. “Non-redundant Subgroup Discovery Using a Closure System”. In: *Proc. ECML PKDD*. Bled, Slovenia, 2009, pp. 179–194. DOI: 10.1007/978-3-642-04180-8_29.

[12] Giorgos Borboudakis and Ioannis Tsamardinos. “Extending greedy feature selection algorithms to multiple solutions”. In: *Data Min. Knowl. Disc.* 35.4 (2021), pp. 1393–1434. DOI: 10.1007/s10618-020-00731-7.

References III

[13] Girish Chandrashekar and Ferat Sahin. "A survey on feature selection methods". In: *Comput. Electr. Eng.* 40.1 (2014), pp. 16–28. DOI: 10.1016/j.compeleceng.2013.11.024.

[14] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C. Tappert. "A Survey of Binary Similarity and Distance Measures". In: *J. Syst. Cybern. Inf.* 8.1 (2010), pp. 43–48. URL: http://www.iiisci.org/Journal/pdv/sci/pdfs/GS315JG.pdf.

[15] Thi-Bich-Hanh Dao and Christel Vrain. "A review on declarative approaches for constrained clustering". In: *Int. J. Approximate Reasoning* 171 (2024). DOI: 10.1016/j.ijar.2024.109135.

[16] Leonardo De Moura and Nikolaj Bjørner. "Z3: An Efficient SMT Solver". In: *Proc. TACAS*. Budapest, Hungary, 2008, pp. 337–340. DOI: 10.1007/978-3-540-78800-3_24.

[17] Daniel Deutch and Nave Frost. "Constraints-based Explanations of Classifications". In: *Proc. ICDE*. Macao, China, 2019, pp. 530–541. DOI: 10.1109/ICDE.2019.00054.

[18] Jonathan Eckstein et al. "The Maximum Box Problem and its Application to Data Analysis". In: *Comput. Optim. Appl.* 23.3 (2002), pp. 285–298. DOI: 10.1023/A:1020546910706.

References IV

[19] Edouard Fouché, Florian Kalinke, and Klemens Böhm. “Efficient subspace search in data streams”. In: *Inf. Syst.* 97 (2021). DOI: 10.1016/j.is.2020.101705.

[20] Jerome H. Friedman and Nicholas I. Fisher. “Bump hunting in high-dimensional data”. In: *Stat. Comput.* 9.2 (1999), pp. 123–143. DOI: 10.1023/A:1008894516817.

[21] José A. Galindo et al. “Automated analysis of feature models: Quo vadis?” In: *Computing* 101.5 (2019), pp. 387–433. DOI: 10.1007/s00607-018-0646-1.

[22] Michael R. Garey and David S. Johnson. *Computers and Intractibility: A Guide to the Theory of NP-Completeness*. 24th ed. W. H. Freeman and Company, 2003. URL: https://www.worldcat.org/title/440655898.

[23] William Christopher Groves. “Toward Automating and Systematizing the Use of Domain Knowledge in Feature Selection”. PhD thesis. University of Minnesota, 2015. URL: https://hdl.handle.net/11299/175444.

[24] D. S. Guru et al. “An alternative framework for univariate filter based feature selection for text categorization”. In: *Pattern Recognit. Lett.* 103 (2018), pp. 23–31. DOI: 10.1016/j.patrec.2017.12.025.

References V

[25] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. "Group Lasso with Overlap and Graph Lasso". In: *Proc. ICML*. 2009, pp. 433–440. DOI: 10.1145/1553374.1553431.

[26] Richard M. Karp. "Reducibility among Combinatorial Problems". In: *Complexity of Computer Computations*. 1st ed. Plenum Press, 1972. Chap. 9, pp. 85–103. DOI: 10.1007/978-1-4684-2001-2_9.

[27] Rami N. Khushaba, Ahmed Al-Ani, and Adel Al-Jumaily. "Feature subset selection using differential evolution and a statistical repair mechanism". In: *Expert Syst. Appl.* 38.9 (2011), pp. 11515–11526. DOI: 10.1016/j.eswa.2011.03.028.

[28] Richard E. Korf. "Objective Functions for Multi-Way Number Partitioning". In: *Proc. SoCS*. Atlanta, GA, USA, 2010, pp. 71–72. DOI: 10.1609/socs.v1i1.18172.

[29] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. "Estimating mutual information". In: *Phys. Rev. E* 69.6 (2004). DOI: 10.1103/PhysRevE.69.066138.

References VI

[30] Vincenzo Lagani et al. “Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets”. In: *J. Stat. Software* 80.7 (2017). DOI: 10.18637/jss.v080.i07.

[31] Nada Lavrač, Peter Flach, and Blaz Zupan. “Rule Evaluation Measures: A Unifying View”. In: *Proc. ILP*. Bled, Slovenia, 1999, pp. 174–185. DOI: 10.1007/3-540-48751-4_17.

[32] Alexander Lawrinenko. “Identical Parallel Machine Scheduling Problems: Structural patterns, bounding techniques and solution procedures”. PhD thesis. Friedrich-Schiller-Universität Jena, 2017. URL: https://nbn-resolving.org/urn:nbn:de:gbv:27-dbt-20170427-0956483.

[33] Matthijs van Leeuwen and Arno Knobbe. “Diverse subgroup set discovery”. In: *Data Min. Knowl. Disc.* 25.2 (2012), pp. 208–242. DOI: 10.1007/s10618-012-0273-y.

[34] Jundong Li et al. “Feature Selection: A Data Perspective”. In: *ACM Comput. Surv.* 50.6 (2017). DOI: 10.1145/3136625.

References VII

Karlsruhe Institute of Technology

[35] Quentin Louveaux and Sébastien Mathieu. “A combinatorial branch-and-bound algorithm for box search”. In: *Discrete Optim.* 13 (2014), pp. 36–48. DOI: 10.1016/j.disopt.2014.05.001.

[36] Tarcísio Lucas, Renato Vimieiro, and Teresa Ludermir. “SSDP+: A Diverse and More Informative Subgroup Discovery Approach for High Dimensional Data”. In: *Proc. CEC*. Rio de Janeiro, Brazil, 2018. DOI: 10.1109/CEC.2018.8477855.

[37] Michael Mampaey et al. “Efficient Algorithms for Finding Richer Subgroup Descriptions in Numeric and Nominal Data”. In: *Proc. ICDM*. Brussels, Belgium, 2012, pp. 499–508. DOI: 10.1109/ICDM.2012.117.

[38] Romain Mathonat et al. “Anytime Subgroup Discovery in High Dimensional Numerical Data”. In: *Proc. DSAA*. Porto, Portugal, 2021. DOI: 10.1109/DSAA53316.2021.9564223.

[39] Brian W. Matthews. “Comparison of the predicted and observed secondary structure of T4 phage lysozyme”. In: *Biochim. Biophys. Acta - Protein Struct.* 405.2 (1975), pp. 442–451. DOI: 10.1016/0005-2795(75)90109-9.

[40] Marvin Meeng and Arno Knobbe. “For real: a thorough look at numeric attributes in subgroup discovery”. In: *Data Min. Knowl. Disc.* 35.1 (2021), pp. 158–212. DOI: 10.1007/s10618-020-00703-x.

References VIII

[41] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. “Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations”. In: *Proc. FAT**. Barcelona, Spain, 2020, pp. 607–617. DOI: 10.1145/3351095.3372850.

[42] Felix Neutatz, Felix Biessmann, and Ziawasch Abedjan. “Enforcing Constraints for Machine Learning Systems via Declarative Feature Selection: An Experimental Study”. In: *Proc. SIGMOD*. Virtual Event, China, 2021, pp. 1345–1358. DOI: 10.1145/3448016.3457295.

[43] Felix Neutatz, Marius Lindauer, and Ziawasch Abedjan. “AutoML in heavily constrained applications”. In: *VLDB J.* (2023). DOI: 10.1007/s00778-023-00820-1.

[44] Randal S. Olson et al. “PMLB: a large benchmark suite for machine learning evaluation and comparison”. In: *BioData Min.* 10 (2017). DOI: 10.1186/s13040-017-0154-4.

[45] Pavel Paclík et al. “On Feature Selection with Measurement Cost and Grouped Features”. In: *Proc. SSPR /SPR*. Windsor, ON, Canada, 2002, pp. 461–469. DOI: 10.1007/3-540-70659-3_48.

References IX

Karlsruhe Institute of Technology

[46] Laurent Perron and Vincent Furnon. *OR-Tools*. Accessed: 2024-06-27. Google, 2024. URL: https://developers.google.com/optimization/.

[47] Jan H. Plasberg and W. Bastiaan Kleijn. "Feature Selection Under a Complexity Constraint". In: *IEEE Trans. Multimedia* 11.3 (2009), pp. 565–571. DOI: 10.1109/TMM.2009.2012944.

[48] Hugo M. Proença et al. "Robust subgroup discovery: Discovering subgroup lists using *MDL*". In: *Data Min. Knowl. Disc.* 36.5 (2022), pp. 1885–1970. DOI: 10.1007/s10618-022-00856-x.

[49] Joseph D. Romano et al. *PMLB v1.0: An open source dataset collection for benchmarking machine learning methods*. arXiv:2012.00058v3 [cs.LG]. 2021. DOI: 10.48550/arXiv.2012.00058.

[50] Arvind Kumar Shekar, Patricia Iglesias Sánchez, and Emmanuel Müller. "Diverse Selection of Feature Subsets for Ensemble Regression". In: *Proc. DaWaK*. Lyon, France, 2017, pp. 259–273. DOI: 10.1007/978-3-319-64283-3_19.

[51] Andreia Silva and Cláudia Antunes. "Constrained pattern mining in the new era". In: *Knowl. Inf. Syst.* 47.3 (2016), pp. 489–516. DOI: 10.1007/s10115-015-0860-5.

References X

[52] Markus Sudmanns et al. “Data-driven exploration and continuum modeling of dislocation networks”. In: *Modell. Simul. Mater. Sci. Eng.* 28.6 (2020). DOI: 10.1088/1361-651x/ab97ef.

[53] Joaquin Vanschoren et al. “OpenML: networked science in machine learning”. In: *ACM SIGKDD Explor. Newsl.* 15.2 (2014), pp. 49–60. DOI: 10.1145/2641190.2641198.

[54] Haiqin Yang et al. “Budget constrained non-monotonic feature selection”. In: *Neural Networks* 71 (2015), pp. 214–224. DOI: 10.1016/j.neunet.2015.08.004.

[55] Ming Yuan and Yi Lin. “Model selection and estimation in regression with grouped variables”. In: *J. R. Stat. Soc. B* 68.1 (2006), pp. 49–67. DOI: 10.1111/j.1467-9868.2005.00532.x.