Fast and Memory-Efficient Discovery of the Top-k Relevant Subgroups in a Reduced Candidate Space

Henrik Grosskreutz & Daniel Paurat

Introduction

Subgroup Discovery and the Theory of Relevance

The Task of Subgroup Discovery

Input:

- examples, characterized by features
- a target class

Output:

- top-*k* subgroup descriptions
  - subgroup that are ***large*** and have a ***high target share***

    e.g. $quality(sd) = n_{sd} \cdot (p_{sd} - p_0)$
  - Subgroup description = conjunction of features

Example:

Approval	Children = yes	Children = no	University	High Income
+		+	+	+
+	+		+	+
+	+			
-		+		+
-	+			
-	+			
-		+		

Fraunhofer IAIS

The Task of Subgroup Discovery

Input:

- examples, characterized by features
- a target class

Output:

- top-*k* subgroup descriptions

1.	High_Inc & University
2.	University
3.	High_Inc
4.	High_Inc & Univ & Children=no
5.	High_Inc & Univ & Children=yes
…	
8.	Children=yes

1. and 2. are equivalent

4. and 5. are irrelevant given 1.

Approval	Children = yes	Children = no	University	High Income
+		+	+	+
+	+		+	+
+	+			
-		+		+
-	+			
-	+			
-		+		

The Theory of Relevance

Def: **Relevance** [Lavrac et al, JLP-99]

- A subgroup is irrelevant if it is dominated
- *s* is dominated by *t* in DB iff.
  - TP(DB,*s*) $\subseteq$ TP(DB,*t*)
  - FP(DB,*s*) $\supseteq$ FP(DB,*t*)

Approval	Children = yes	Children = no	University	High Income
+		+	+	+
+	+		+	+
+	+			
-		+		+
-	+			
-	+			
-		+		

Example:

"HighInc&Univ&Child=no" is dominated by **"HighInc&Univ"**

Relevant top-*k* Subgroup Discovery

Lavrac & Gamberger: Relevancy in constraint-based subgroup discovery, 2005

Input:

- a set of examples characterized by features
- a target class

Output:

- the *k* highest-quality ***relevant*** subgroup descriptions

Description	Classic sd	Closed sd	Relevant sd
High_Inc & University	1st	1st	1st
University	2nd		
High_Inc	3rd	2nd	
Children=yes & High_Inc & University	4th	3rd	
Children=yes & High_Inc	5th		
Children=no & High_Inc & University	6th	4th	
Children=no & High_Inc	7th		
Children=yes	8th	5th	2nd

Existing Approaches

… and their limitations

Pruning-based Approaches to Relevant SD

e.g. Lemmerich & Atzmueller: Fast discovery of relevant subgroup patterns, FLAIRS 2010

Idea:

- Traverse the space of subgroup descriptions, e.g. using DFS
  - Keep track of the *k* best subgroup visited
- Apply pruning
  - Use quality of the *k*-best subgroup ("$\boldsymbol{\theta_k}$") as minimum quality threshold
  - Prune branches whose quality can be derived to be below θ_k
- **Local relevance check**
  - Whenever a new high-quality subgroup is visited, check dominance between the *k*+1 best subgroups

➔ **Output consists of relevant subgroups**

Pruning-based Approaches: Limitations

Pruning based on a local relevance check can miss relevant subgroups

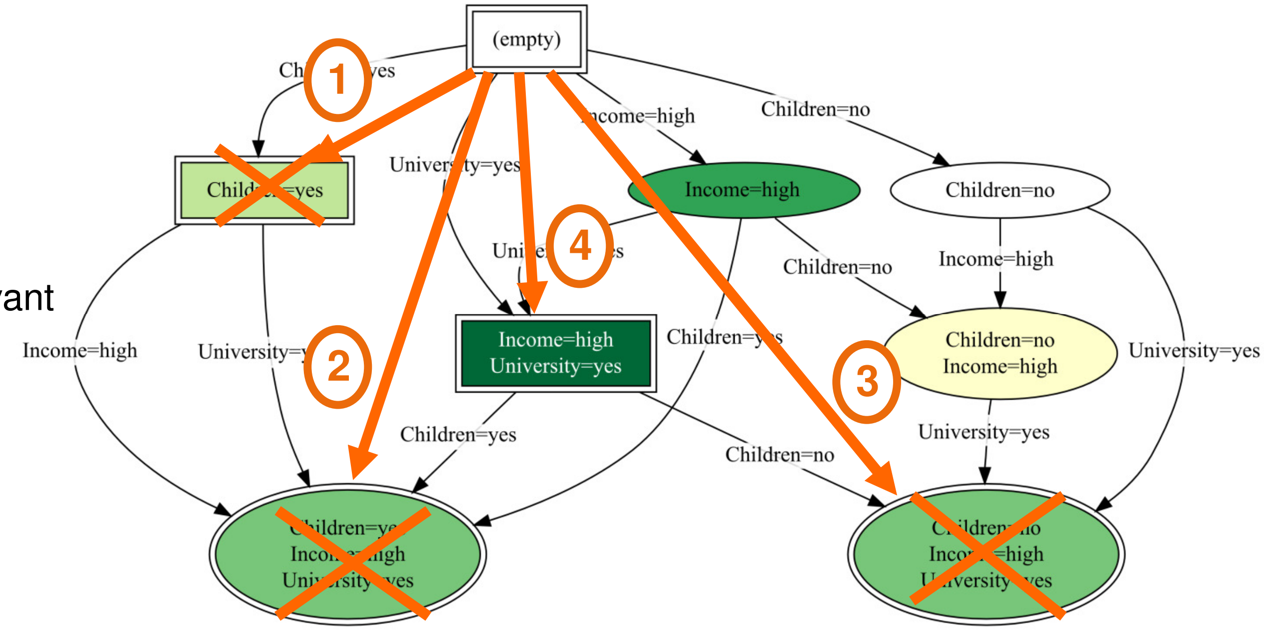
➔ **Problem: local relevance check is not an exact test**

Example:

Intensity ~ quality

Rectangular shape = relevant

Rel-SD with k=2



Fraunhofer IAIS

Approach based on the Closed-on-the-positives

Garriga et al.: Closed sets for labeled data, JML-08

- Proposition: A subgroup description *sd* is relevant iff.
  - It is **closed on the positives**
  - There is no cpos generalization $s_g \subseteq sd$ with same support in the negatives
- Approach:
  - Collect all closed-on-the-positives
  - Remove irrelevant subgroups in a post-processing step

➔ **Advantages: exact**
 # cpos can be exponentially smaller than # closed / all sd

➔ **Problems: huge memory requirements; no pruning**

Summary of the existing Approaches

- Pruning-based approach:
  - doesn't guarantee exact results
- C-pos approach:
  - infeasible for large number of c-pos
  - no pruning

A new Approach

… based on iterative deepening and an efficient relevance check

Efficient relevant subgroup discovery

An efficient algorithm should

1. only consider the *closed-on-the-positive* subgroups
2. avoid high memory requirements
3. apply pruning based on θ_k
   - Requires an exact relevance check at visiting time

An *O(k)* Relevance Check

Proposition:

For many popular quality functions*, relevance of a closed-on-the-positive *sd* can be checked based only the ***higher-quality*** relevant generalizations

$G^* = \{s_g \subseteq sd \mid s_g$ is relevant and has ***higher quality*** than sd$\}$

Hence:

If we are only interested in relevance of ***subgroups with quality > θ_k,***
and we visit the ***cpos*** in a **general-to-specific** fashion
then relevance can be checked using ***only the top-k subgroups visited***

→ Memory requirements: *k* subgroups, instead of $O(2^{length(sd)})$

* : in particular, for $q(sd) = n^a (p-p_0)$, with $0 \leq a \leq 1$

Fraunhofer IAIS

The New Algorithm ID-Rsd

Idea

- Perform an iterative deepening
- Only keep track of the best *k* subgroups visited
- Perform relevance check by comparing with the *k* best subgroups

Properties:

- *Exact solution*
- *Memory requirements*: O(k) subgroups (+ Iterative deepening DFS)
- *Max. number of nodes visited is* $O(|\mathbb{C}_p| \cdot n)$, where *n* ~ number of features
- *Allows pruning based on* θ_k

Depending on shape of search space

Fraunhofer IAIS

Comparison with existing Approaches

Algorithm	Memory	Runtime	Pruning
Classic SD	$O(n^2 + kn)$	$O(\lvert\mathbb{S}\rvert\, n\, m)$	Yes
Closed SD	$O(n^2 + kn)$	$O(\lvert\mathbb{C}\rvert\, n^2\, m)$	Yes
RelSD_Classic	No exact result, otherwise like above		
RelSD_Cpos	$\Omega(\lvert\mathbb{C}_p\rvert\, n)$	$\Omega(\lvert\mathbb{C}_p\rvert\, n^2\, m)$	No
ID-Rsd	$O(n^2 + kn)$	$O(\lvert\mathbb{C}_p\rvert\, (n^3\, m + n^2\, k))$	Yes

+

No worse than classic/closed

+

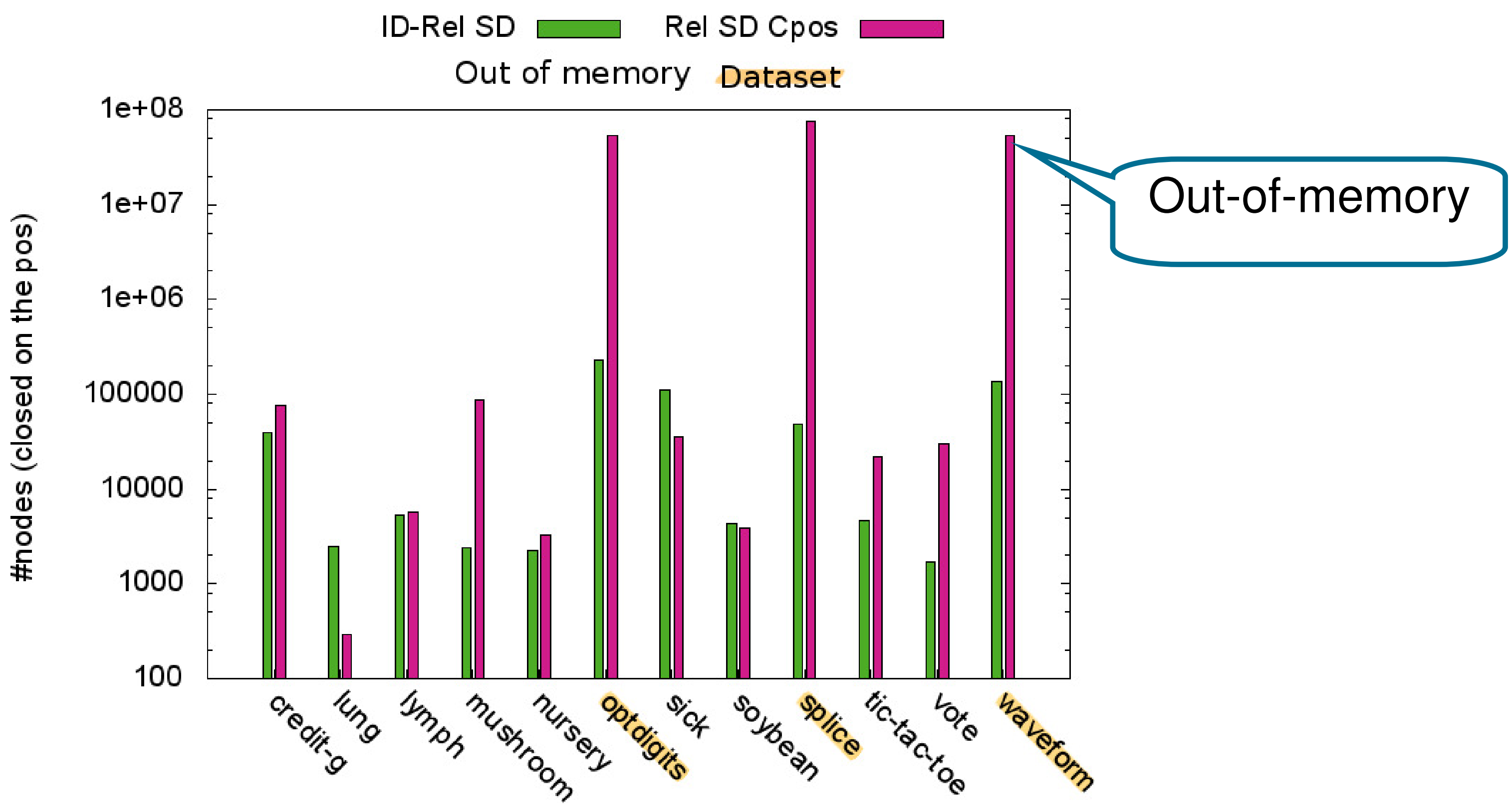
$\mathbb{S}$ ~ Subgroups (all)
$\mathbb{C}$ ~ Closed subgroups
$\mathbb{C}_p$ ~ Closed on the positives

n ~ # features
m ~ # records

Empirical Evaluation

Fraunhofer IAIS

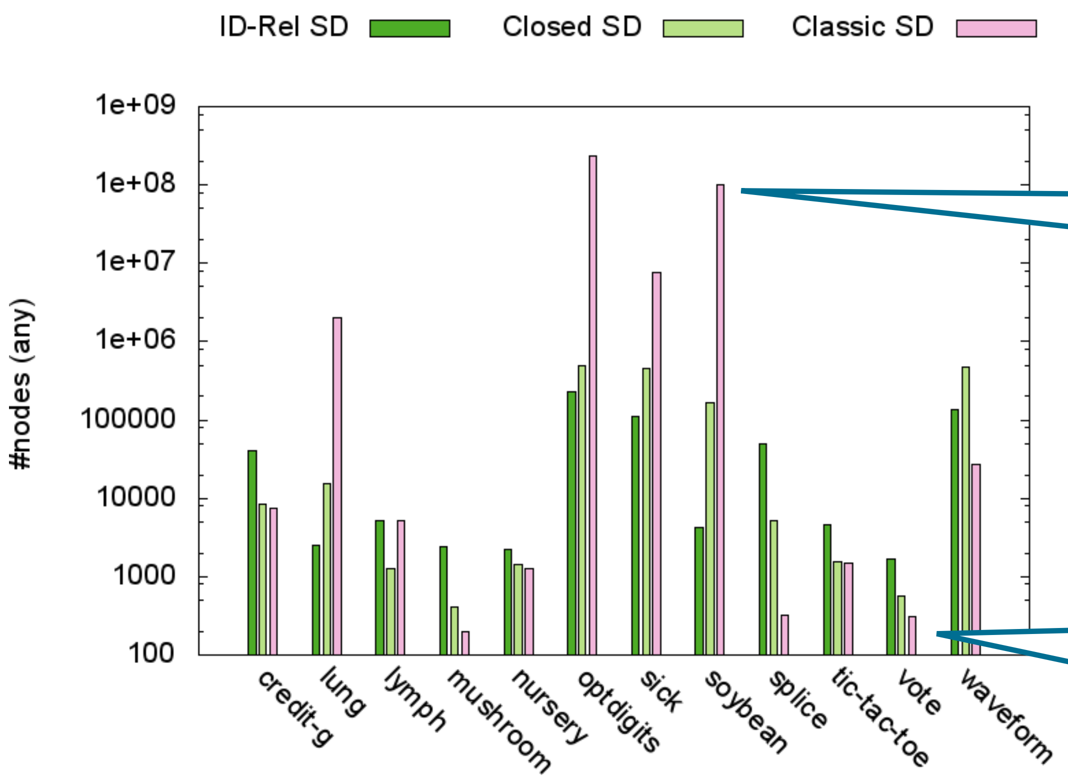
Comparison with the Closed-on-the-positives Approach



BinTest, k=10

Overall: lower memory requirements, lower number of nodes & much faster

Comparison with Classic & Closed SD Approaches



Overhead can be exponential

Caused i.a. by different θ_k

	ID-Rsd	ClassicSD	ClosedSD	Cpos-Rsd
Total runtime	118 sec	2717 sec	286 sec	?

Summary

Summary

- Relevant SD yields more valuable patterns than classic SD
- ID-Rsd
  - First exact Rsd approach with polynomial memory requirements
  - Much faster than C-pos approach
  - Competitive with exhaustive classic/closed SD approaches

Thank you very much for your attention!