On Subgroup Discovery in Numerical Domains

Henrik Grosskreutz and Stefan Rüping

Fraunhofer IAIS

Fraunhofer
IAIS

Overview

Introduction: **Subgroup discovery** in **numerical** domains

- Motivation and definition
- Problems

New algorithm:

- Based on a new **pruning scheme**
- Empirical evaluation

Subgroup Discovery – A Local Pattern Discovery Task

Task: Find descriptions of subgroups of the overall population that are both

- Large
- Unusual label

Class attribute

Subgroup description: **Conjunction** of **attribute-value** pairs

Example:
Profession= Teacher & Sex=M
→ Cost=High

Example: Employee dataset

Stress	**Profession**	**Sex**
High	Teacher	M
High	Baker	F
High	Scientist	F
Low	Baker	M
Low	Teacher	F
High	Teacher	M

The Quality of a Subgroup

Quality Functions: $n^a(p - p_0)$

- n: Size of the subgroup extension
- p: target share in subgroup
- p_0: overall target share
- $0 \leq a \leq 1$: Constant
  - a=1: "Piatetsky-Shapiro"/"WRAC"
  - a=0.5: Binomialtest

Ex.: Profession= Teacher & Sex=M

- P-S Quality = 2 (1 – 4/6)

Stress	Profession	Sex
High	Teacher	M
High	Baker	F
High	Scientist	F
Low	Baker	M
Low	Teacher	F
High	Teacher	M

Subgroup Discovery in *Numerical (or Ordinal) Domains*

In many domains, the attributes are not nominal but numeric (or ordinal)

Subgroup descriptions: Conjunctions of **attribute-interval** pairs

Examples

- BMI $\in$ [26,30] and BP $\in$ [160,180]
- BMI $\in$ [26,30]

Example: Medical dataset

Vascular disease	**Blood Pressure (systolic)**	**Body Mass Index**
Yes	160	30
No	160	22
No	120	22
Yes	180	20
No	120	30
Yes	170	26

How to deal with numerical attributes?

Standard approach: (**Entropy) Discretization**

- (A) Replace every numeric attribute by ***one*** nominal attribute ranging over ***non-overlapping*** intervals

Result

D(X') = {[1-2],[3-4],[5,6]}

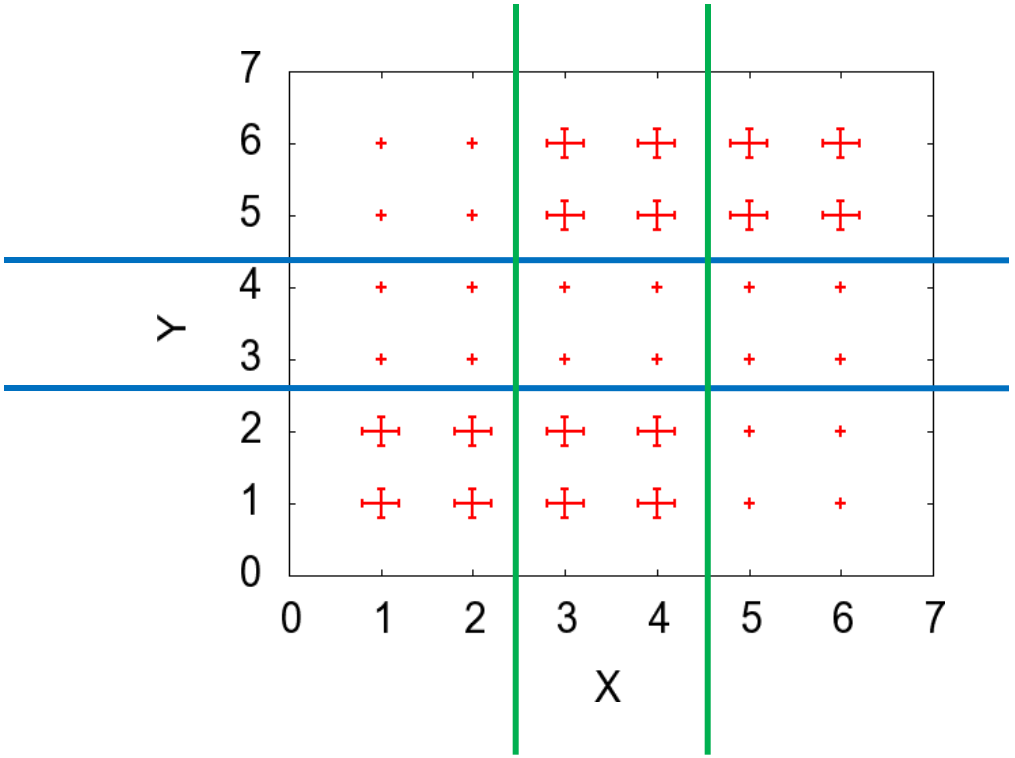
Problems

Expected result should include

- X ∈ [1-4] and Y ∈ [1-2]

But we only obtain:

- X = [1-2] and Y = [1-2] ,
- X = [3-4]



Fraunhofer IAIS

How to deal with numerical attributes (ii)?

Standard approach: **Entropy Discretization**

- (B) Replace every numeric attribute by a ***set*** of binominal attributes, i.e. use ***overlapping intervals***

D(Y') = {[1-2],[3-4],[5,6],[1-4],[1-6],[3-6]}

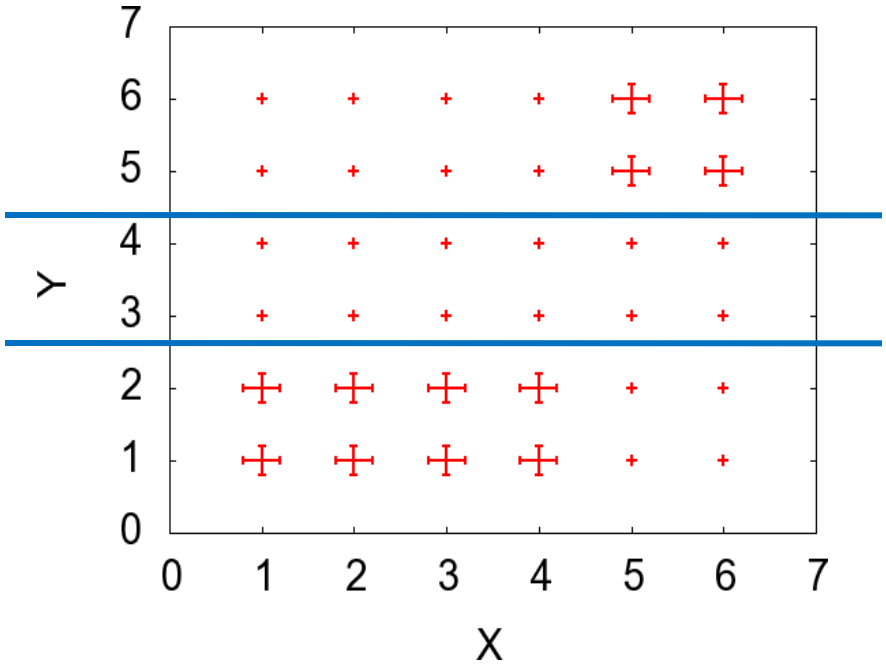
D(X') = Ø

Problem:

Expected solution should include:

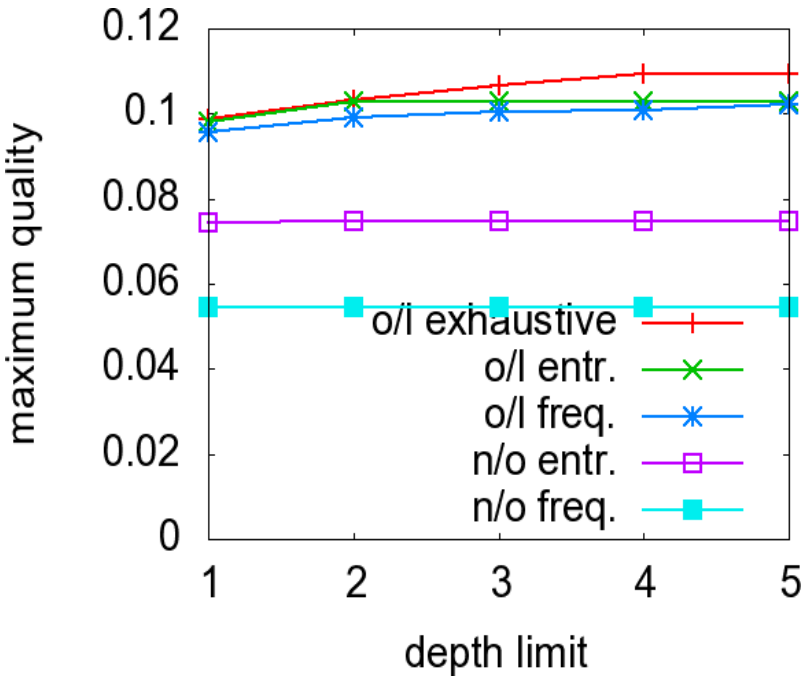
- X ∈ [1-4] and Y ∈ [1-2]

But the obtained result will not contain *any* constraint on X

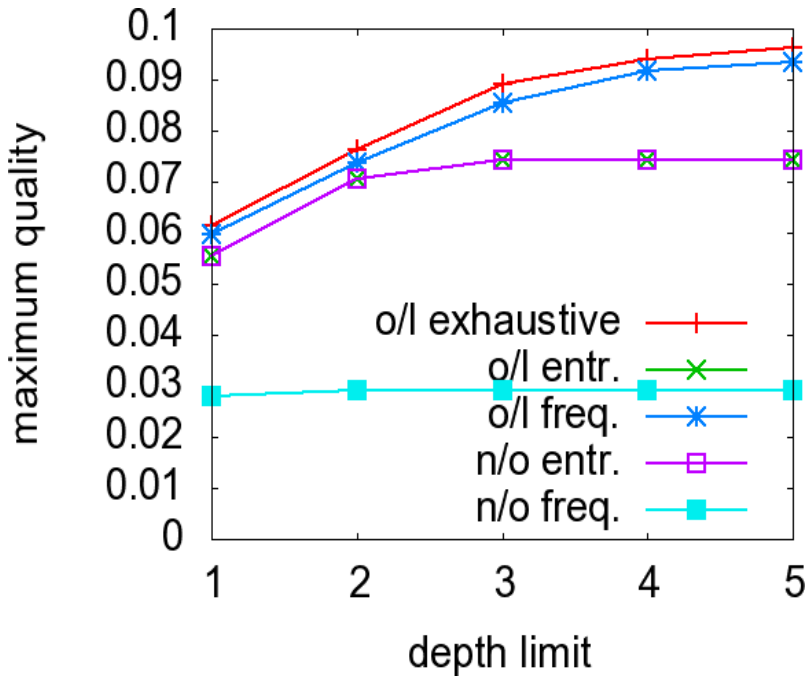


Discretization Strategies on Benchmark Datasets: Quality of the *Best* Subgroup

'diabetes' dataset



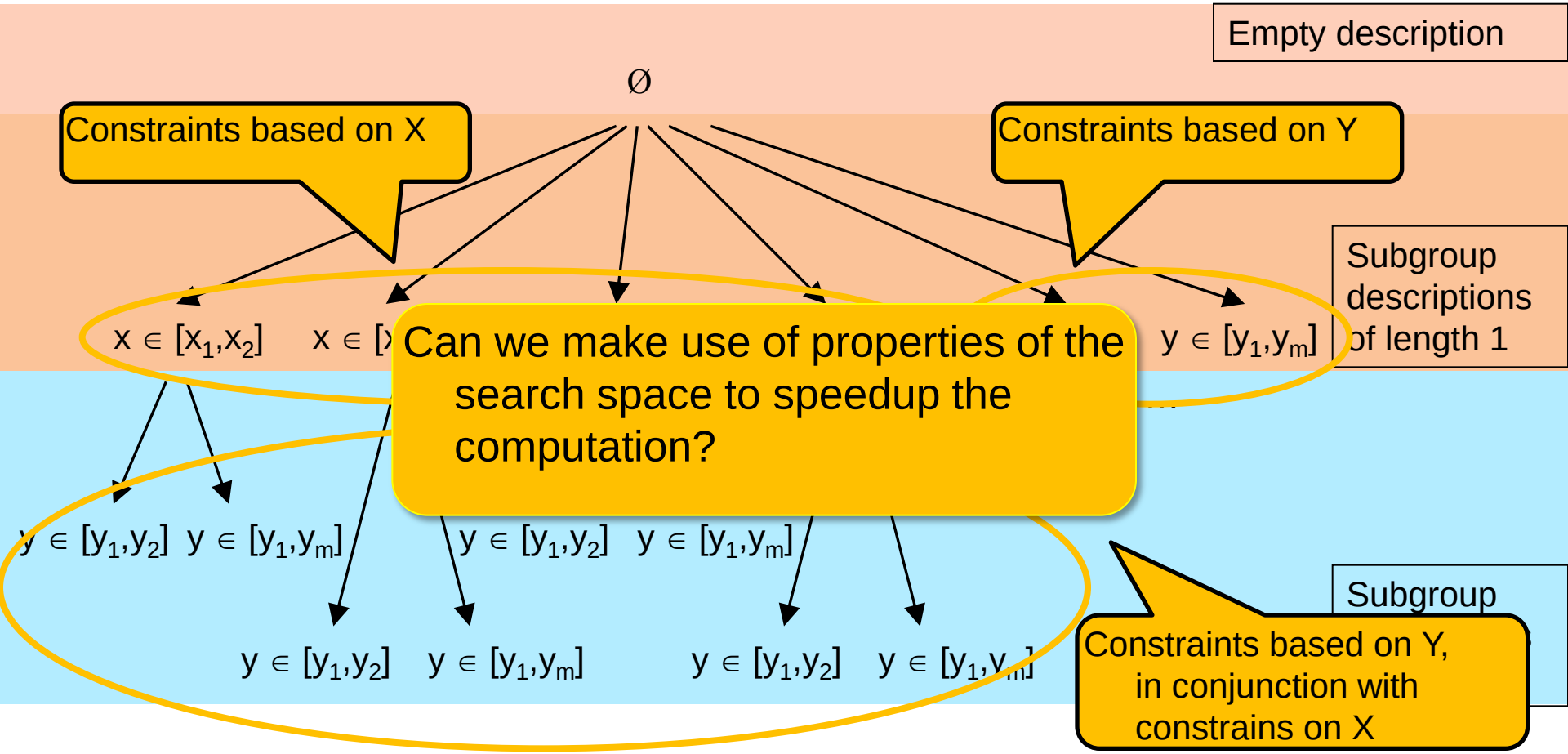
'yeast ' dataset



To find optimal subgroup descriptions, we have to consider **attribute-interval** constraints over arbitrary intervals

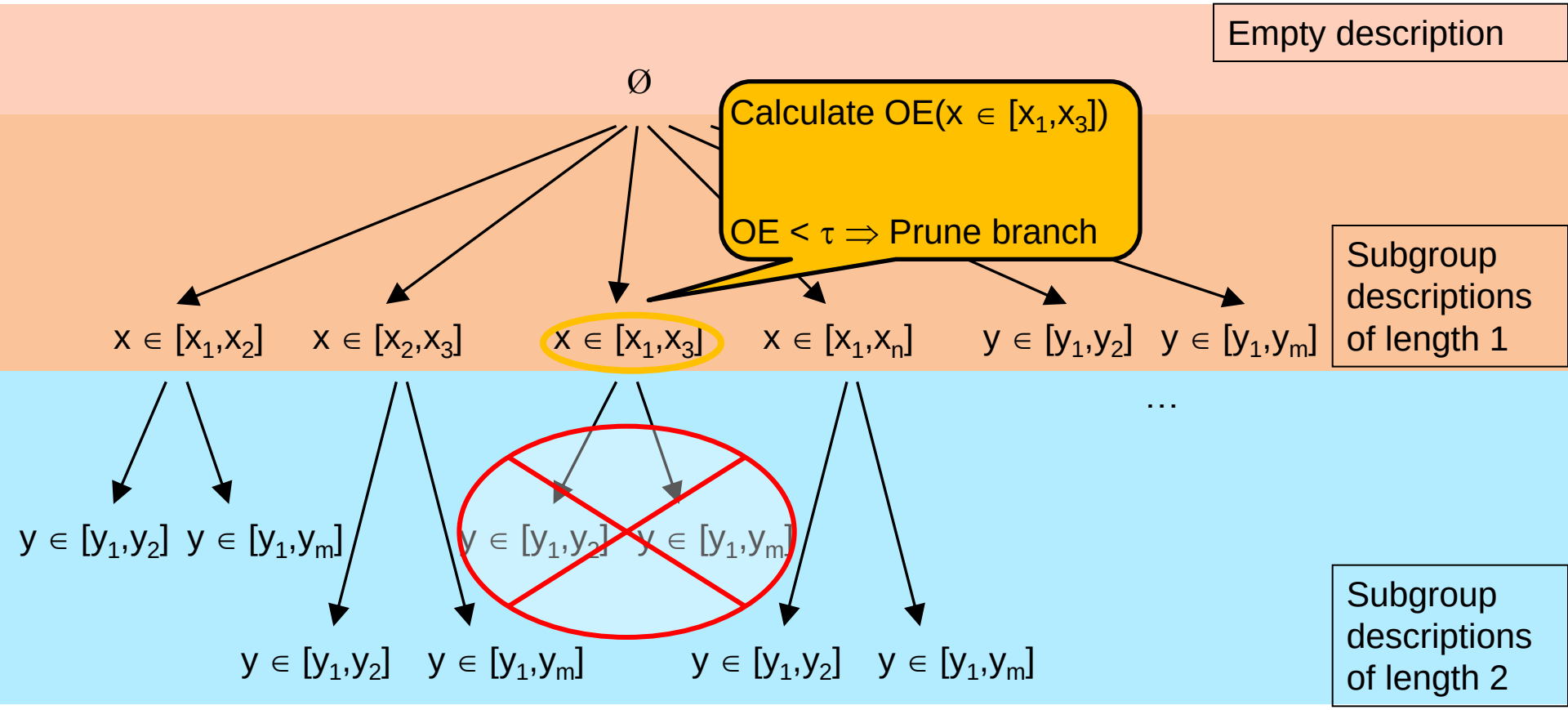
How can we improve the performance if **overlapping** intervals are considered?

Subgroup Discovery as Search in the Space of Subgroup Descriptions

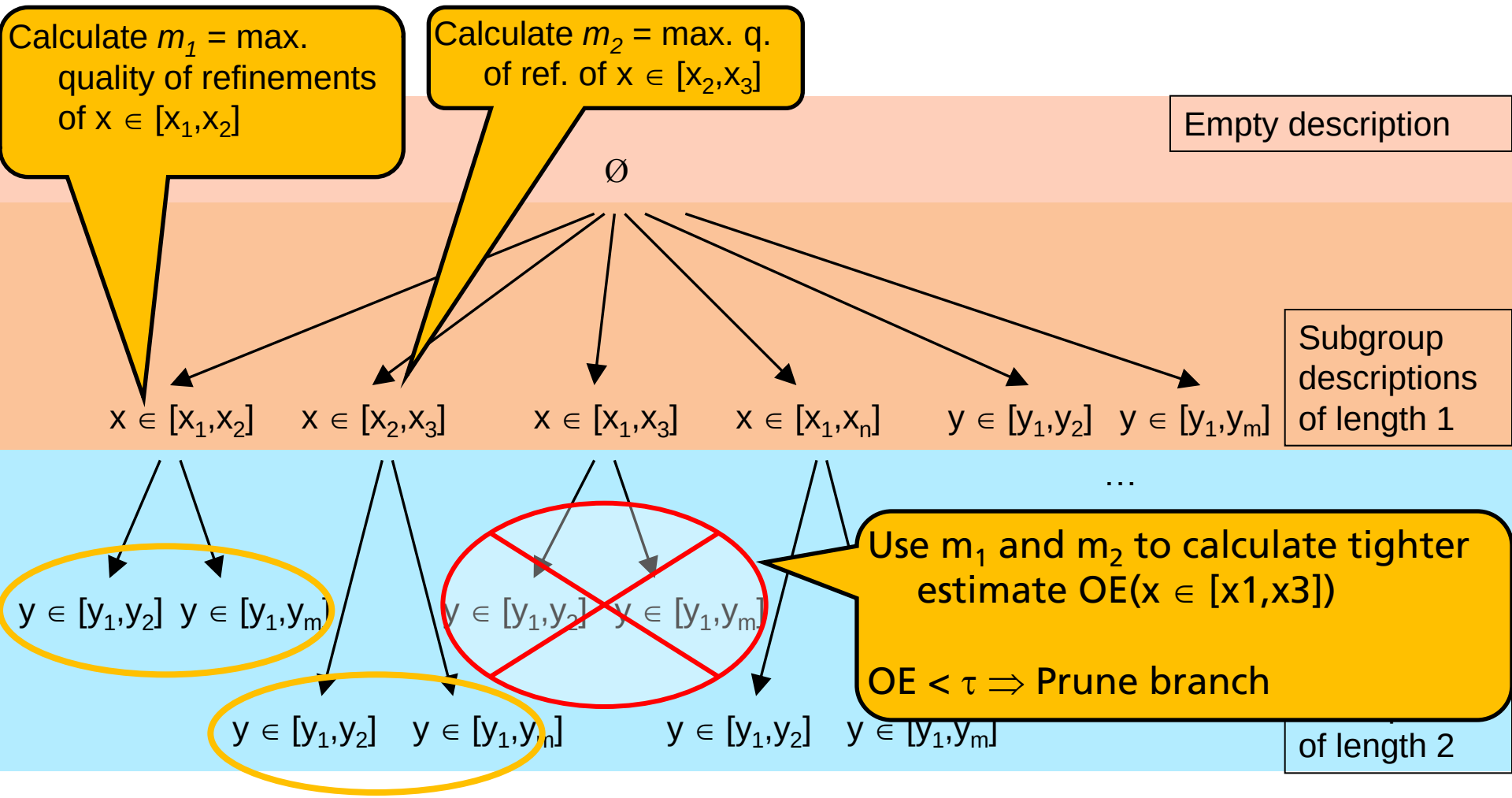


Fraunhofer IAIS

Standard Approach:
DFS with Optimistic Estimate Pruning (Horizontal Pruning)



New Approach: “Horizontal Pruning”



Lemma

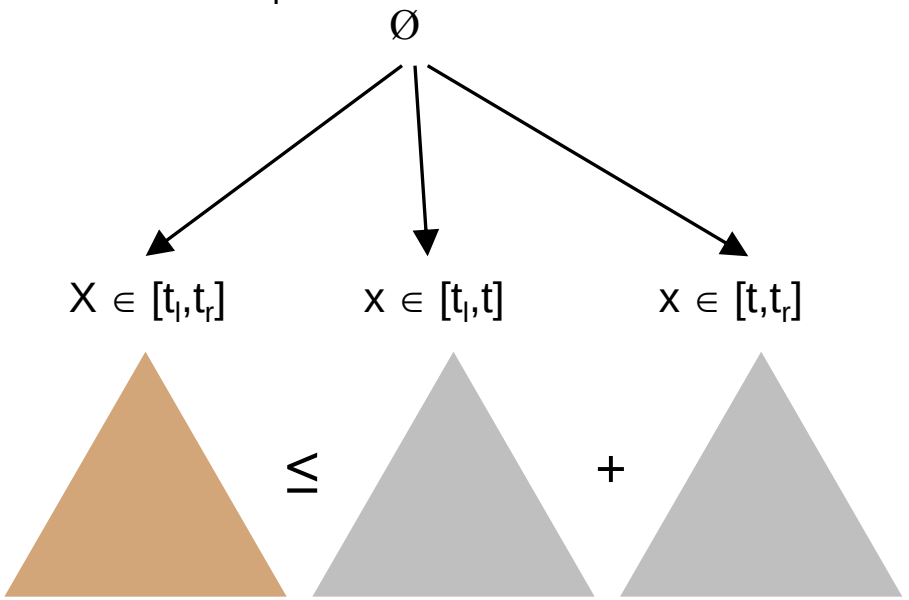
Let t_l and t_r be two split points. The quality $n^a(p - p_0)$, $(a \leq 1)$ of every refinement of

$$sd \wedge X \in [t_l, t_r]$$

is bound by the *sum* of the maximum of the qualities of all refinements of

$$sd \wedge X \in [t_l, t] \quad \text{and} \quad sd \wedge X \in [t_l, t]$$

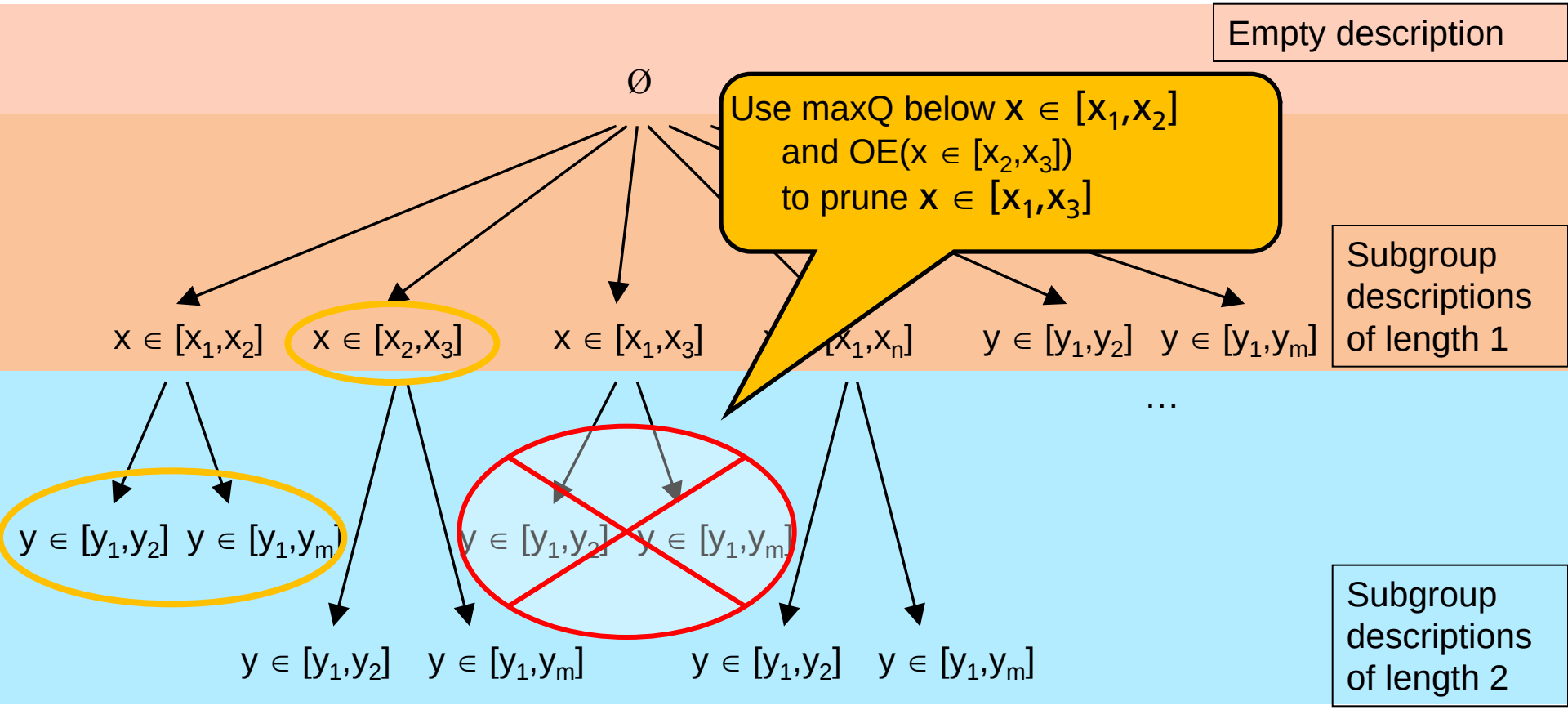
for every t in $[t_l, t_r]$



This property

- Also holds if a depth limit is considered
- Also holds if an arbitrary set of candidate split points is considered

Combining Exact Bounds and Classic Optimistic Estimates



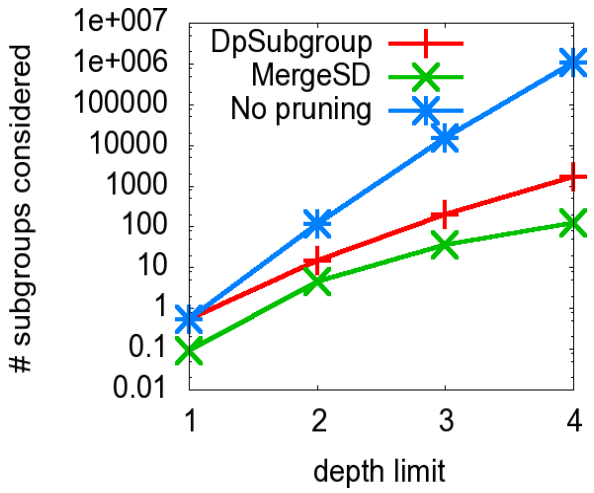
A New Algorithm for SD in Numerical Domains

Main Idea:

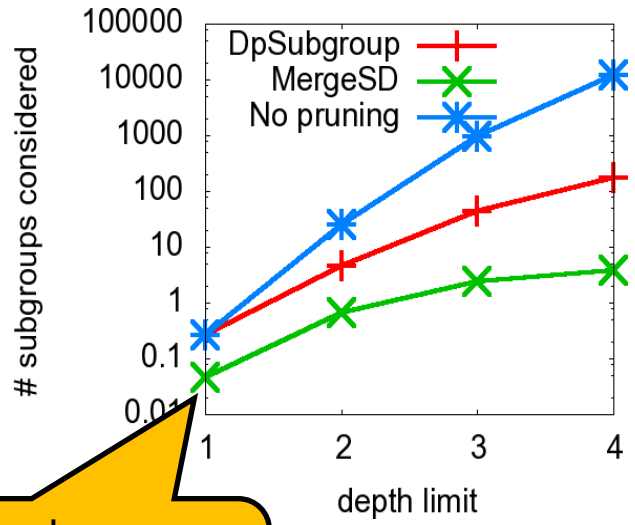
- DFS in space of subgroup descriptions
- Uses optimistic estimates to initialize the Bound tables
- When new bound maxQ for subgroups involving $X \in [t_l,t_r]$ becomes available
  - Set Bound[l,r] to maxQ
  - $\forall$ l', r'. $l' \leq l, r \leq r'$
    Bound [l', r'] = Bound [l', r]+maxQ+Bound [l, r'].
- Heuristik: Never consider an interval if its subinterval have not yet been considered

Experimental Results

Number of nodes considered
when using frequency discretization with 10 bins



diabetes

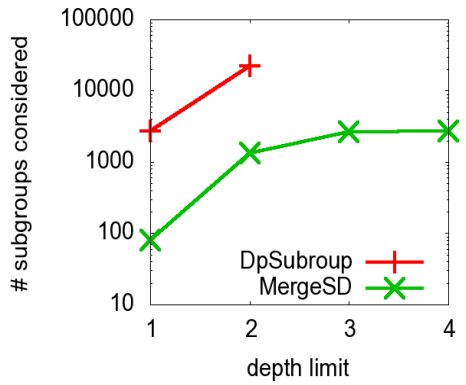


transfusion

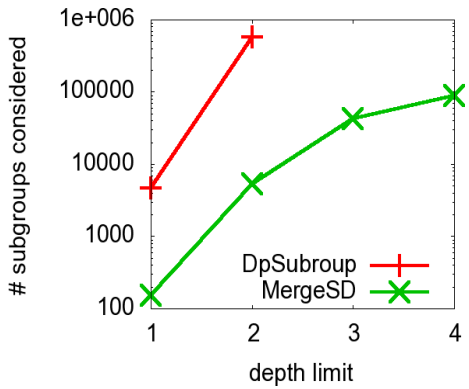
Pruning affects subgroup descriptions of length 1!

Fraunhofer IAIS

Experimental Results (ii): Speedup



mamography



transfusion

Dataset	freq. discr. (10 splits)	exhaustive
breast-w	6	27
diabetes	35	?
fraud	534	?
letter	107	872
mammography	3	139
spambase	125	?
transfusion	14	6312
yeast	92	?

Speedup @ maximum length 2

Summary

- Classical **subgroup discovery approaches…**
- … have problems in **numeric domains.**
- New **“horizontal” pruning scheme** speeds up computation

Open Issues

- Good Sets of Subgroups (Iterated Weighted Covering?)
- Mixed Domains
- Further Speedups
- …

Thank you for your attention!