FSSD – A Fast and Efficient Algorithm for Subgroup Set Discovery

Authors: Adnene and Aimene Belfodil, **Anes Bendimerad**, Philippe Lamarre, Céline Robardet, Mehdi Kaytoue, Marc Plantevit

Several Icons are made by authors from www.flaticon.com

KDD and Subgroup Discovery

" KDD [ed., Knowledge Discovery in Databases] is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data "

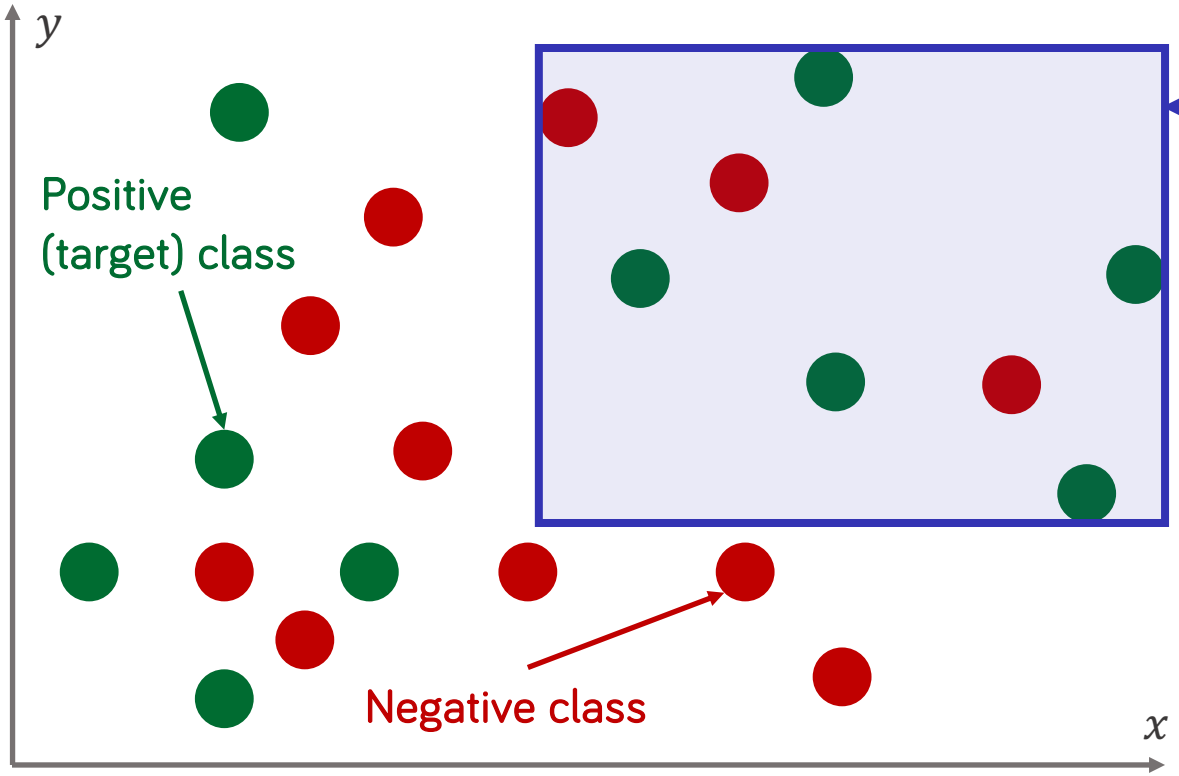
Fayyad U., Piatetsky-Shapiro G., Smyth P. [AAAI, 1996]

" ... A particularly important subclass of knowledge discovery tasks is the discovery of interesting subgroups in populations ... "

S. Wrobel [PKDD, 1997]

Siebes. Data Surveying: Foundations of an Inductive Query Language. KDD (1995)
S. Wrobel. An Algorithm for Multi-relational Discovery of Subgroups. PKDD (1997)
W. Klösgen. Explora: Multipattern and Multistrategy Discovery Assistant. Advances in Knowledge Discovery and Data Mining (1996)

Discriminative Subgroup Discovery



Subgroup. A subset of instances selected by:

$$6 \leq x \leq 16 \; \boldsymbol{and} \; 3 \leq y \leq 9$$

Interestingness. e.g. the informedness:

True positive rate (tpr) – False positive rate (fpr)

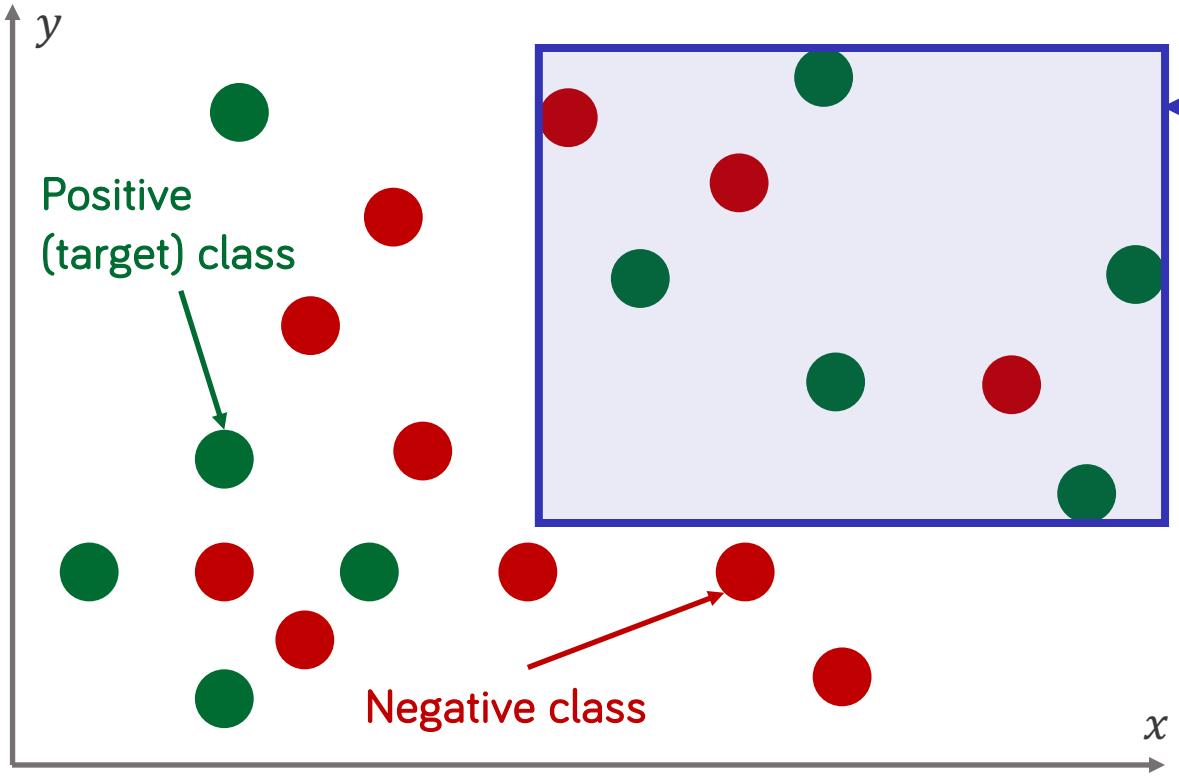
$$\frac{5}{10} - \frac{3}{10} = \frac{2}{10}$$

Subgroup positively correlated to the target class!

Data. Set of **labeled** instances described by attributes.

P. K. Novak, N. Lavrac, G. I. Webb..Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining JMLR 2009

Discriminative Subgroup Discovery



Data. Set of **labeled** instances described by attributes.

Subgroup. Any subset of instances selected by some query in a **description language**.

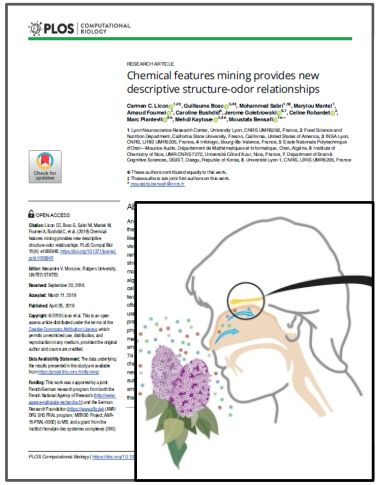
e.g. All conjunctions of intervals

Interestingness. Any measure **increasing** with the **true positive rate** and **decreasing** with **false positive rate***

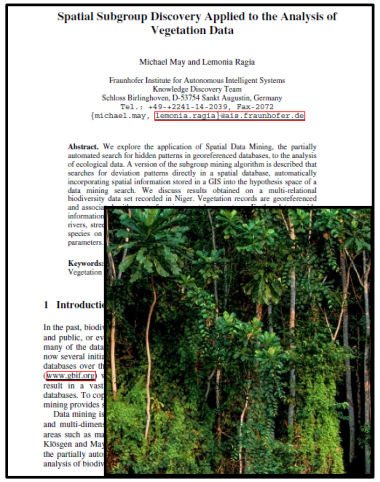
More than **30** measures: WRAcc, Binomial test, Klösgen measure, Linear correlation, Cohen's κ, G-measure, F-measure, m-estimate, Discriminatively, Accuracy, Odds ratio, Growth rate, Lift, Brin's factor, Zhang measure ...

*This property is stronger than the Piatetsky-Shapiro like (**P2**,**P3**)

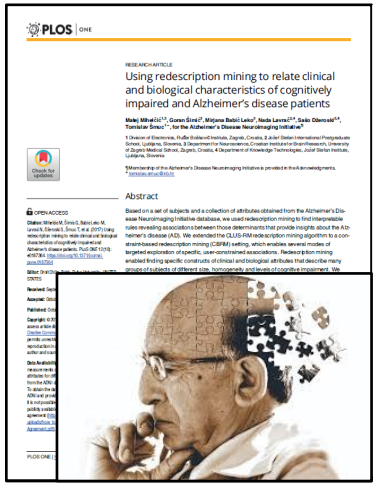
Discriminative Subgroup Discovery - Examples



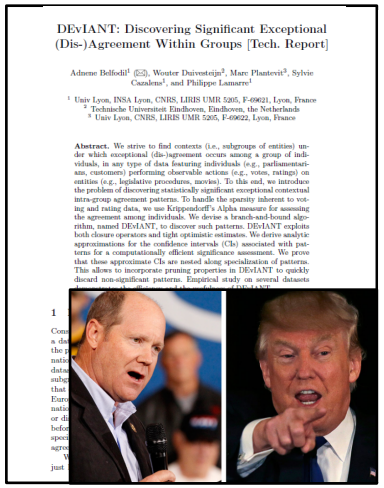
Find relationships between the physicochemical properties of odorant molecules and their olfactory qualities



Find dependencies of a plant species on other plant species and environmental parameters.

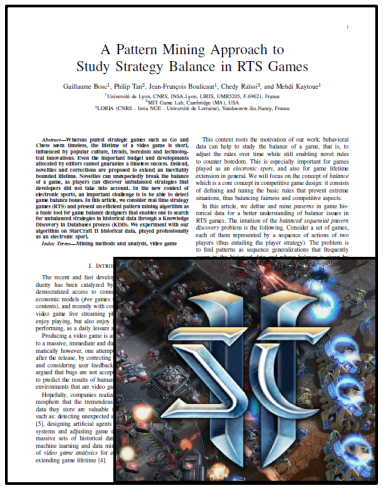


Find relationship between biological and clinical attributes of cognitively impaired and Alzheimer's disease patients



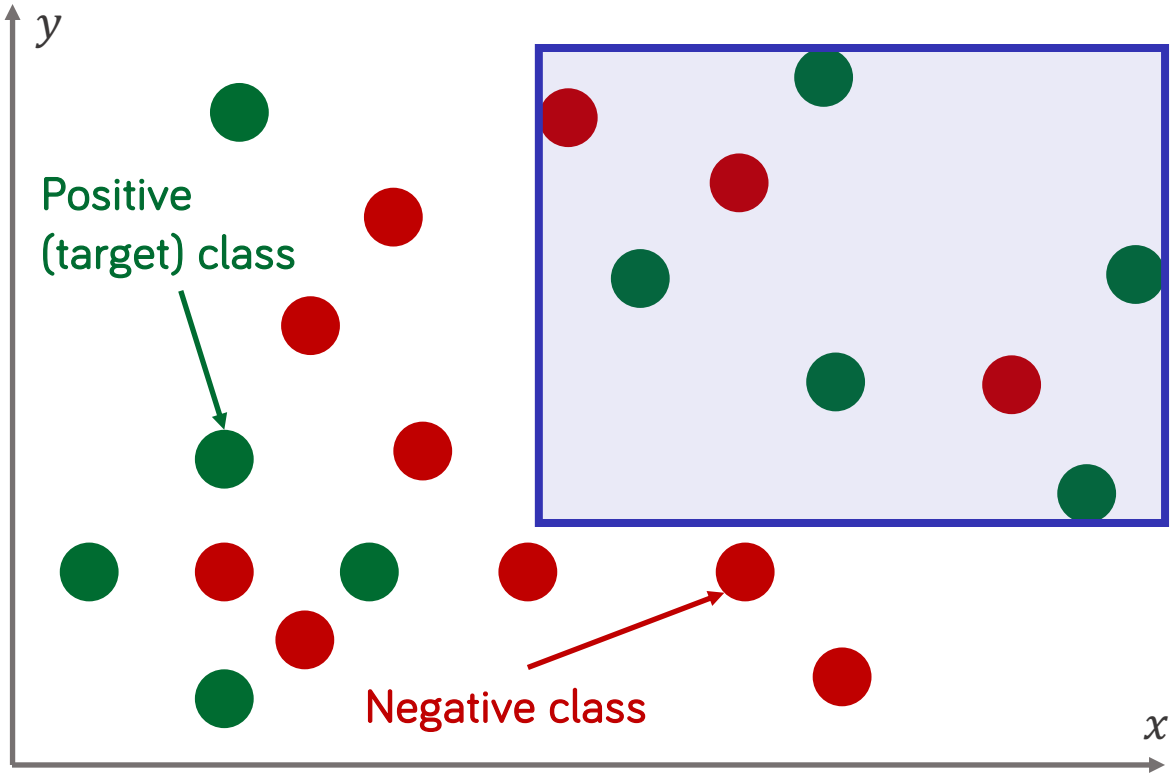
Find exceptional disagreements within groups of individuals.
e.g., Pro and Anti Trump within Republicans in 2016

Capture the periodic structure present in the data.
e.g., Customer buying behavior, Sacha daily activities.

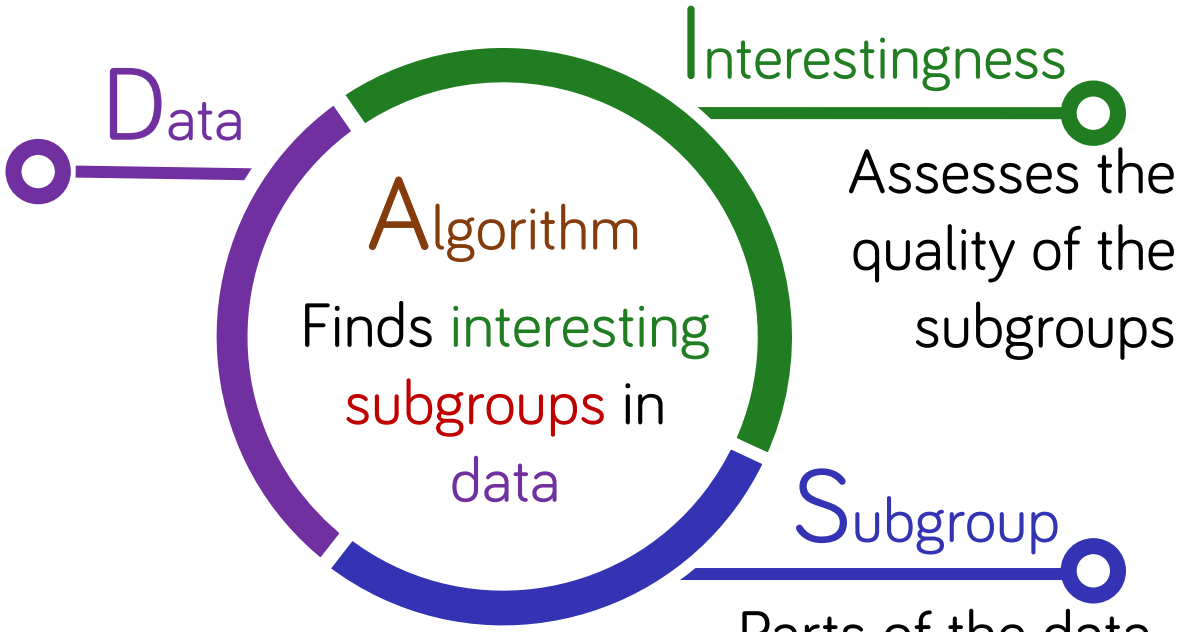


Find unbalanced strategies in historical data of real-time strategy games.
e.g., the unbalanced bunker rush strategy

Discriminative Subgroup Discovery

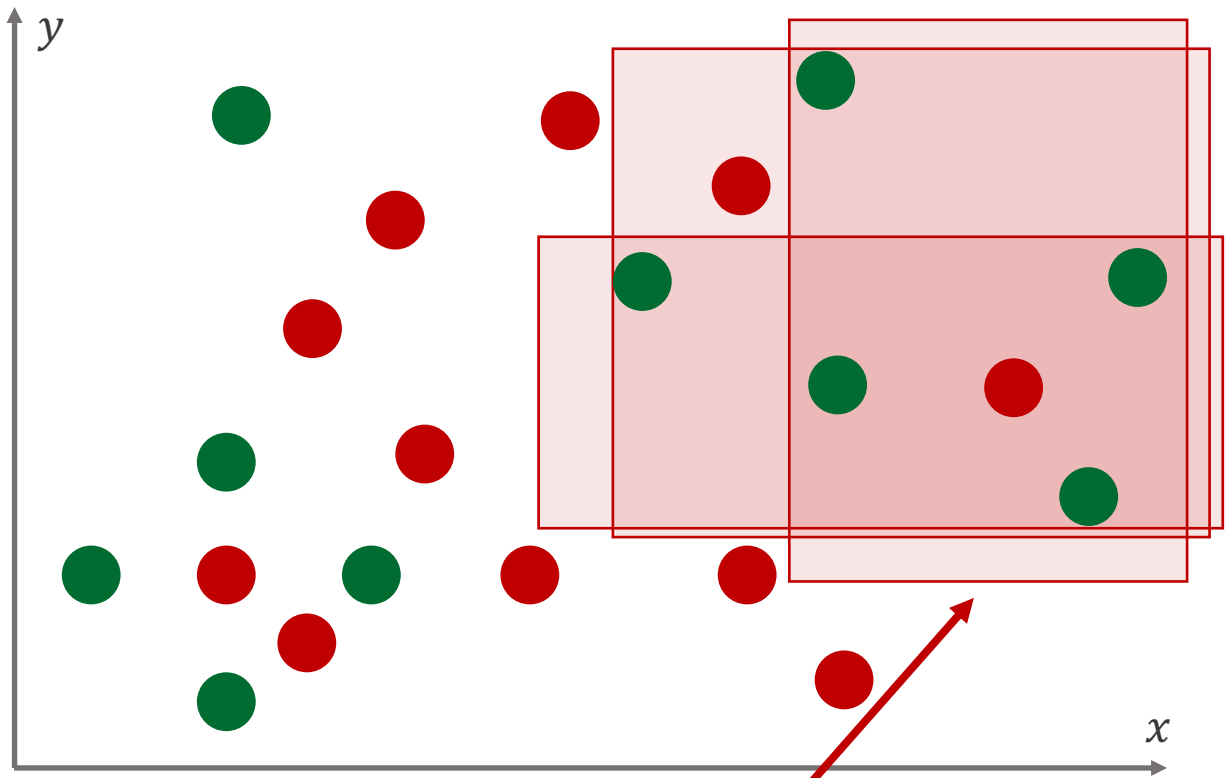


Data. Set of **labeled** instances described by attributes.



P. K. Novak, N. Lavrac, G. I. Webb..Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining JMLR 2009

Problem Statement



Top-3 (relevant/closed) subgroups
w.r.t. Informedness (or WRAcc)

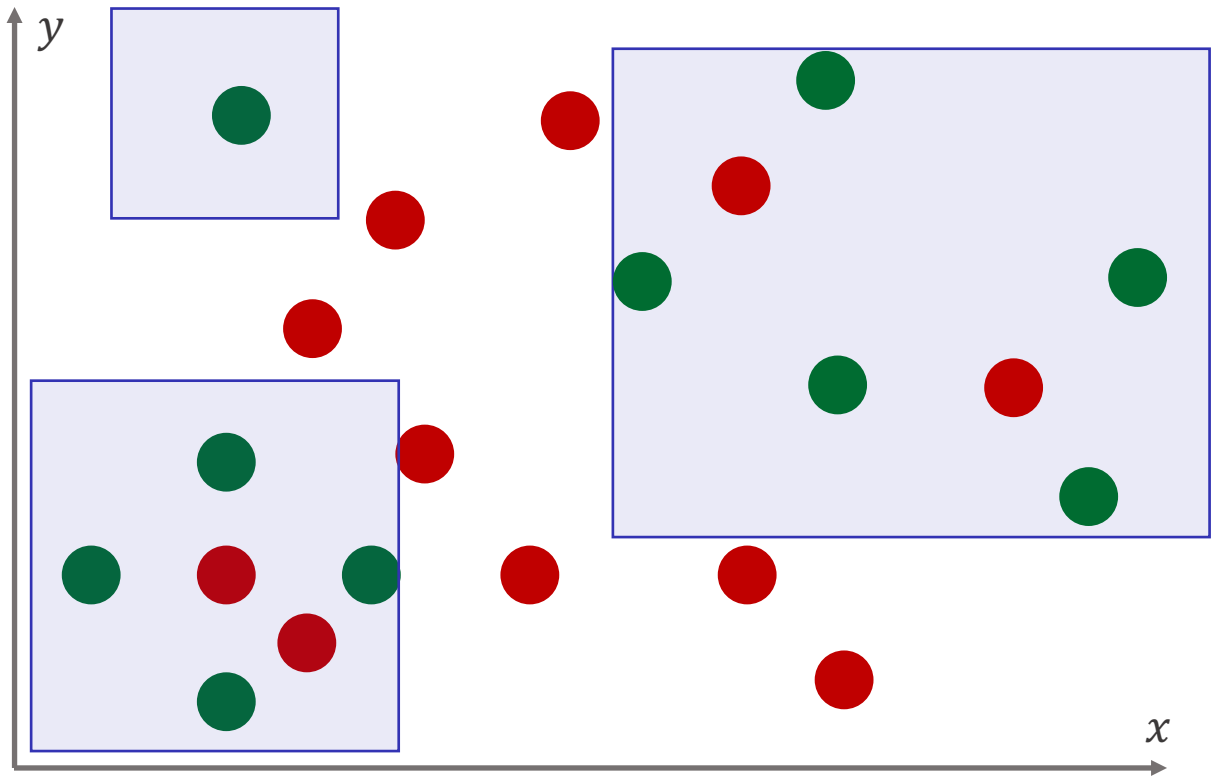
Find a set of at most k interesting subgroups in a labeled dataset

Solution 1: Find the Top-k interesting subgroups

Problem: Non diverse, Top-k subgroups tend to cover the SAME region.

F. Lemmerich, M. Rohlfs, M. Artzmüller. Fast Discovery of Relevant Subgroup Patterns. Flairs 2010
M. Riondato, F. Vandin. MiSoSouP: Mining Interesting Subgroups with Sampling and Pseudodimension. KDD 2018

Problem Statement



*The subgroups are not necessarily disjoint

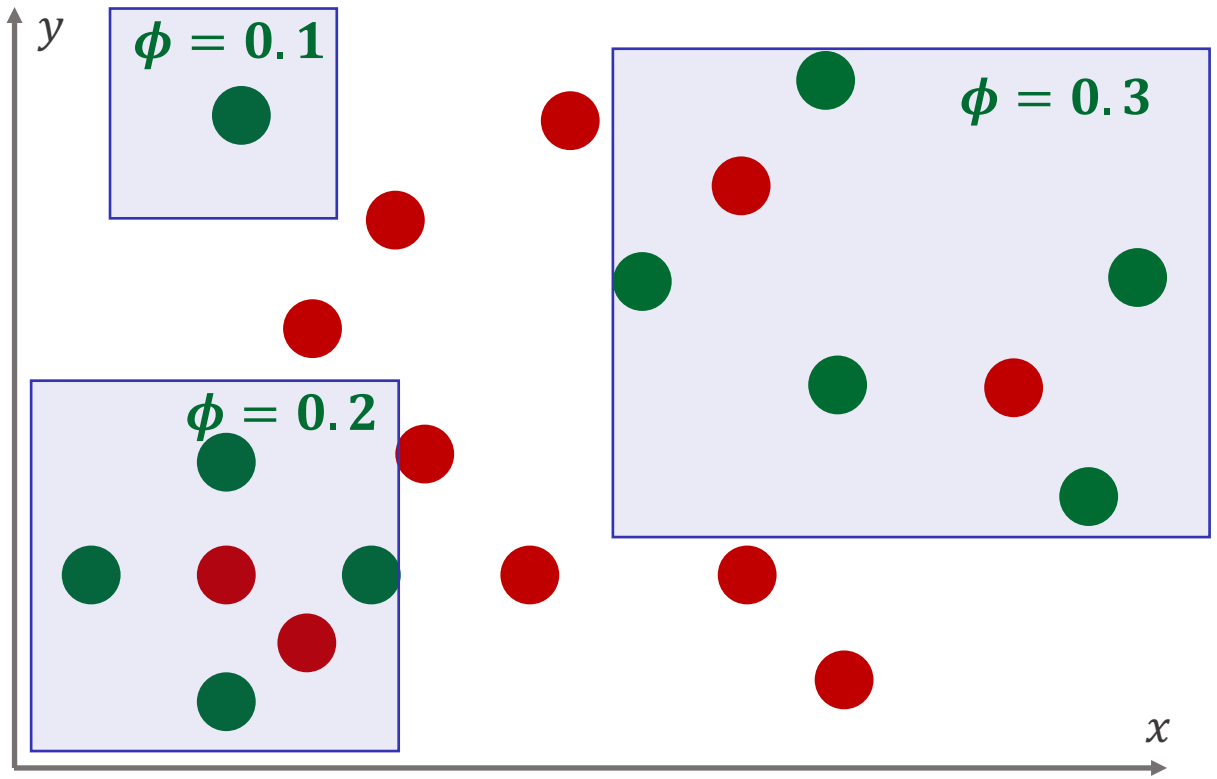
Find a **Diverse** set of at most k interesting subgroups in a labeled dataset

Solution 2: Add a NEW measure to evaluate the redundancy of the subgroup set and use it:

- in some exact algorithms,
- on-the-fly or at the end of some heuristic algorithms.

B. Bringmann, A. Zimmermann. The Chosen few: On Identifying Valuable Patterns. ICDM 2007.
T. Guns, S. Nijssen, L. De Raedt. K-pattern Set Mining under Constraints.. TKDE 2013.
M. van Leeuwen, A. Knobbe. Diverse Subgroup Set Discovery. DAMI 2012
G. Bosc, J. F. Boulicault, C. Raïssi, M. Kaytoue. Anytime discovery of a diverse set of patterns with Monte Carlo tree search. DAMI 2018

Our Problem Statement



$$\phi\ (\textit{subgroup set}) = 0.6$$

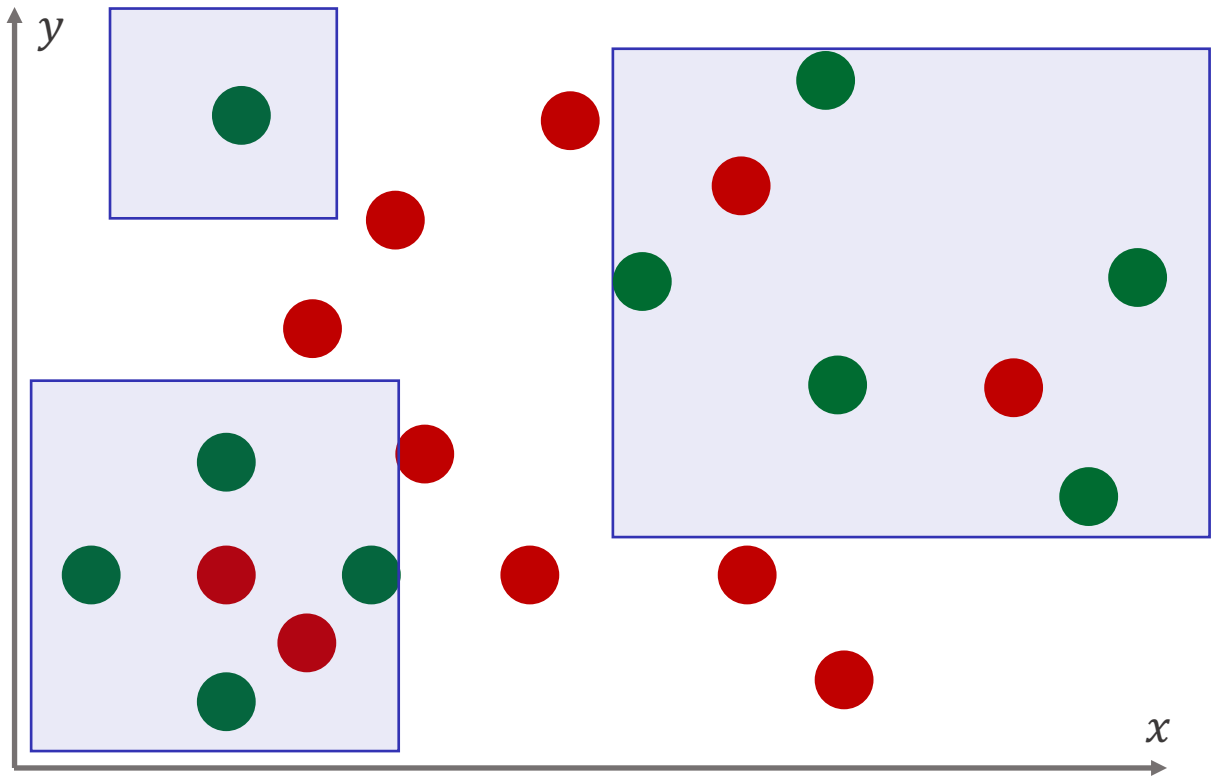
Find a **Diverse** set of at most k interesting subgroups in a labeled dataset

Idea: use the SAME measure ϕ by evaluating the quality of the union (i.e., disjunction):

$$\phi\ (\textit{subgroup set}\ \mathcal{S}) = \phi\left(\bigcup_{s \in \mathcal{S}} s\right)$$

This evaluates diversity implicitly!

Our Problem Statement and Contribution



Find a **Diverse** set of at most k interesting subgroups in a labeled dataset

$$\phi\ (\textbf{subgroup set}\ \mathcal{S}) = \phi\left(\bigcup_{s \in \mathcal{S}} s\right)$$

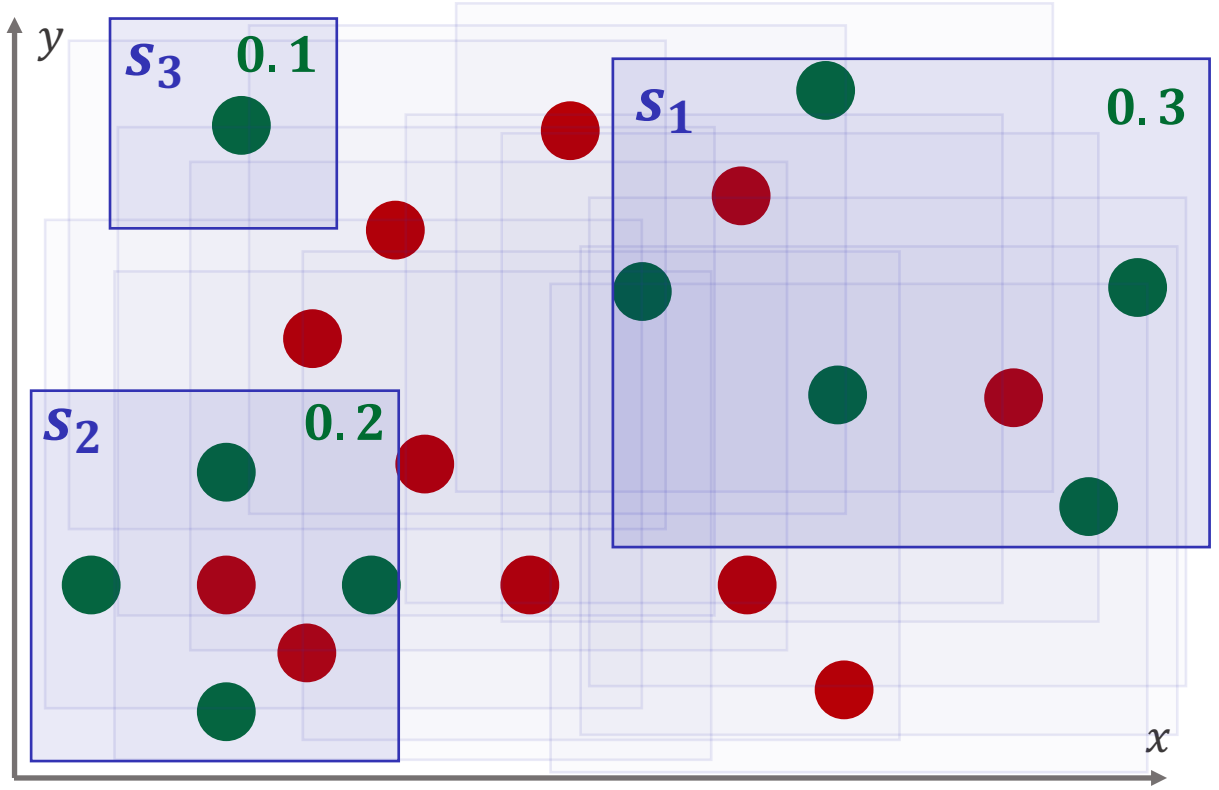
Contribution: Propose an efficient algorithm (FSSD) that maximizes this measure **greedily**.

Problem: Finding the best k-subgroup set w.r.t. this measure is a **NP-hard problem** (proof in the paper)

Outline

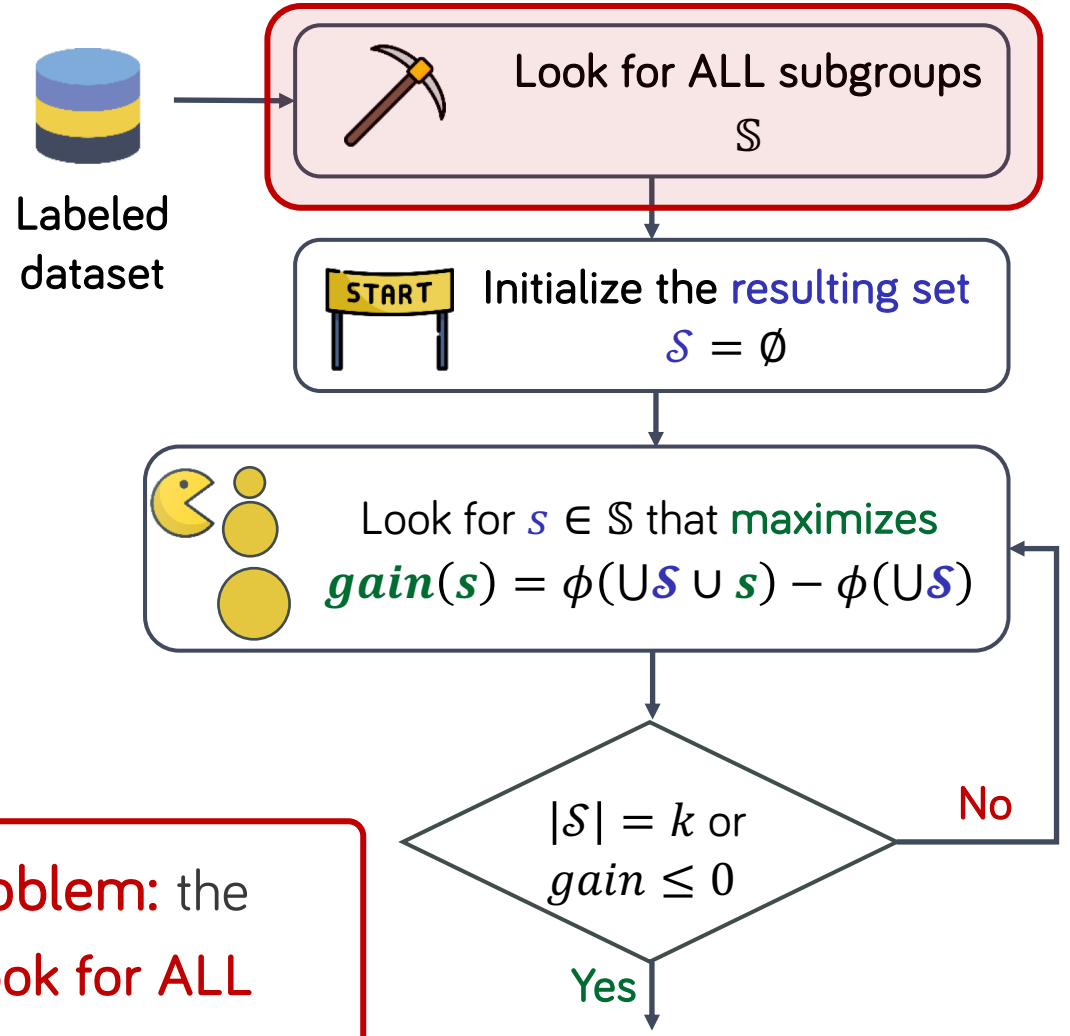
Outline

Greedy Scheme – A naïve approach



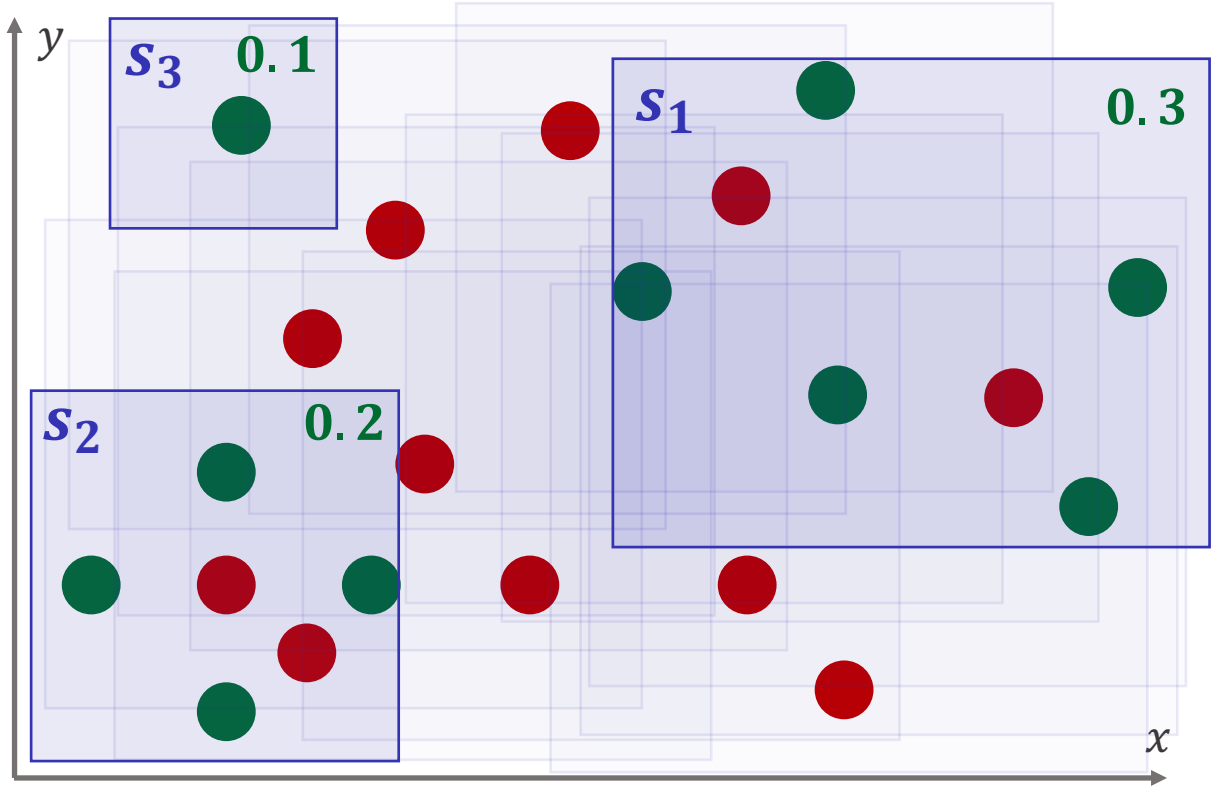
Output $\mathcal{S} = \{ s_1 , s_2 , s_3 \}$

Parameters.
- Subgroup set size $k = 3$
- Measure $\phi = tpr - fpr$



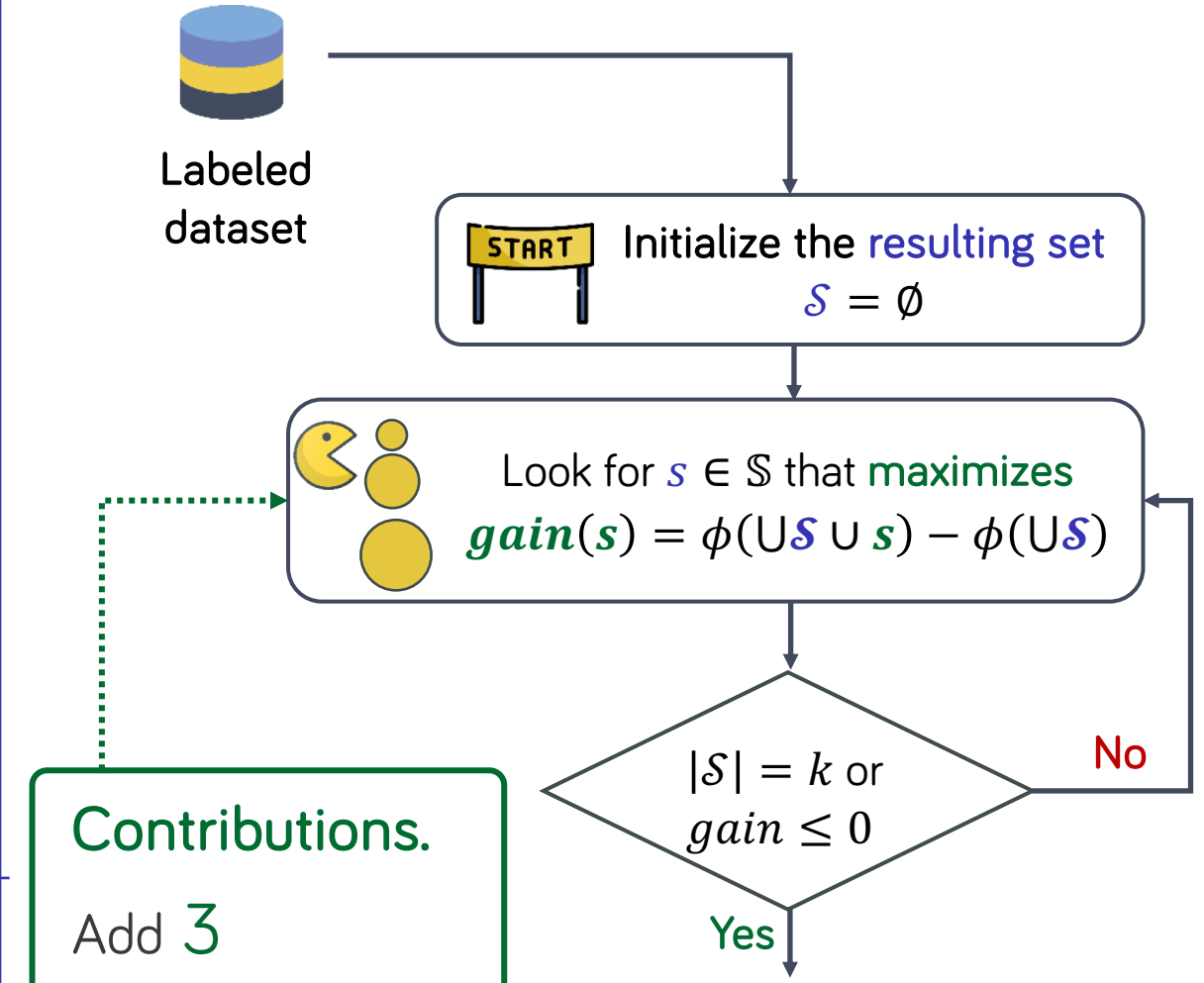
Problem: the "Look for ALL subgroups" step.

Greedy Scheme – Algorithm FSSD



Output $\mathcal{S} = \{ s_1 , s_2 , s_3 \}$

Parameters.
- Subgroup set size $k = 3$
- Measure $\phi = tpr - fpr$



Contributions.

Add 3 Optimizations

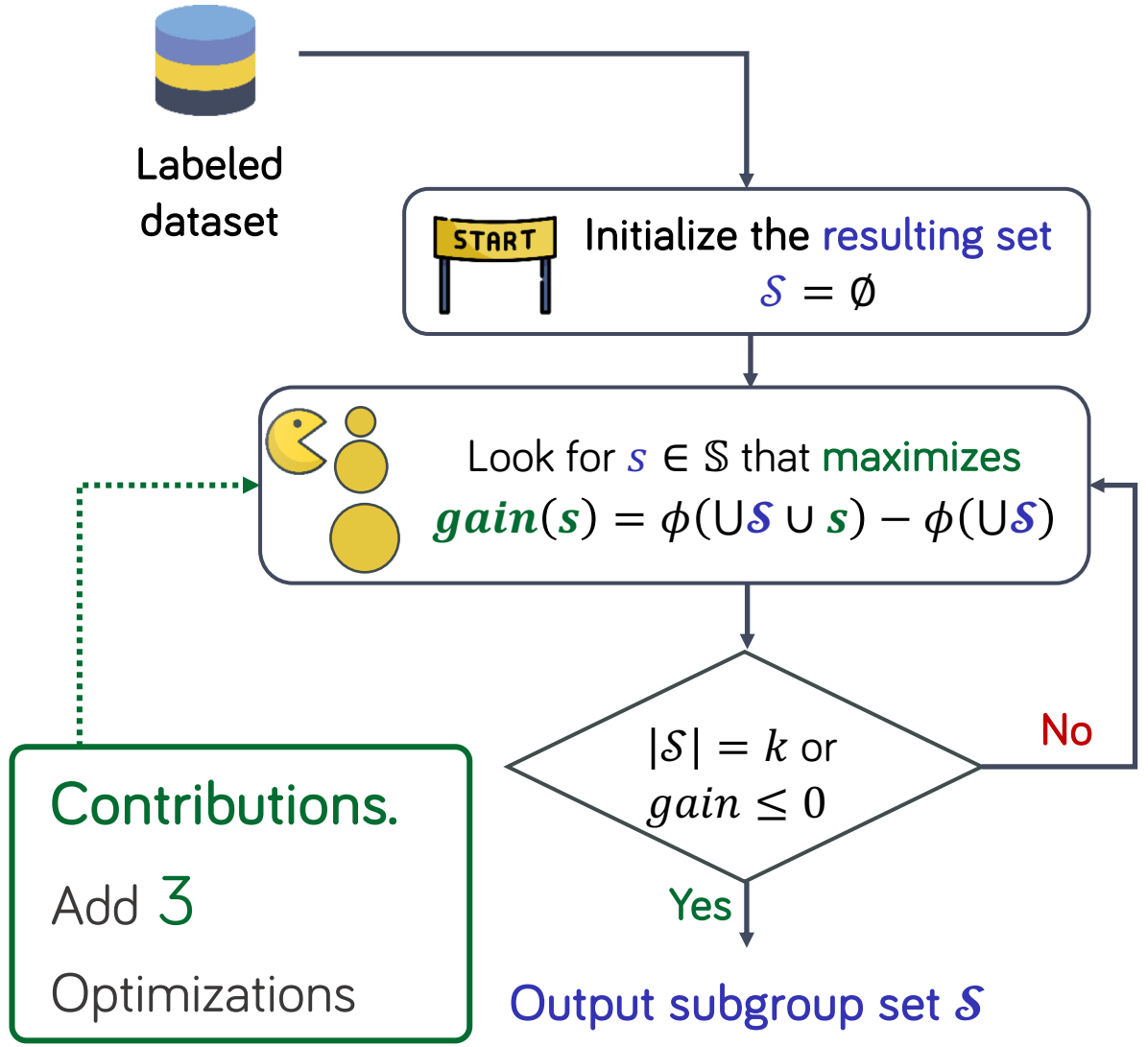
Greedy Scheme – Algorithm FSSD

1. Ignore unpromising branches:

Tight optimistic estimates on the $gain$

- If ϕ is increasing with the tpr and decreasing with the fpr then the $gain$ has the same properties!
- Branch-and-bound using the tight Optimistic estimates: For or all subgroups $t \subseteq s$, we have $gain(t) \leq gain(s^{+})$

  Positive instances in current subgroup s



Contributions.

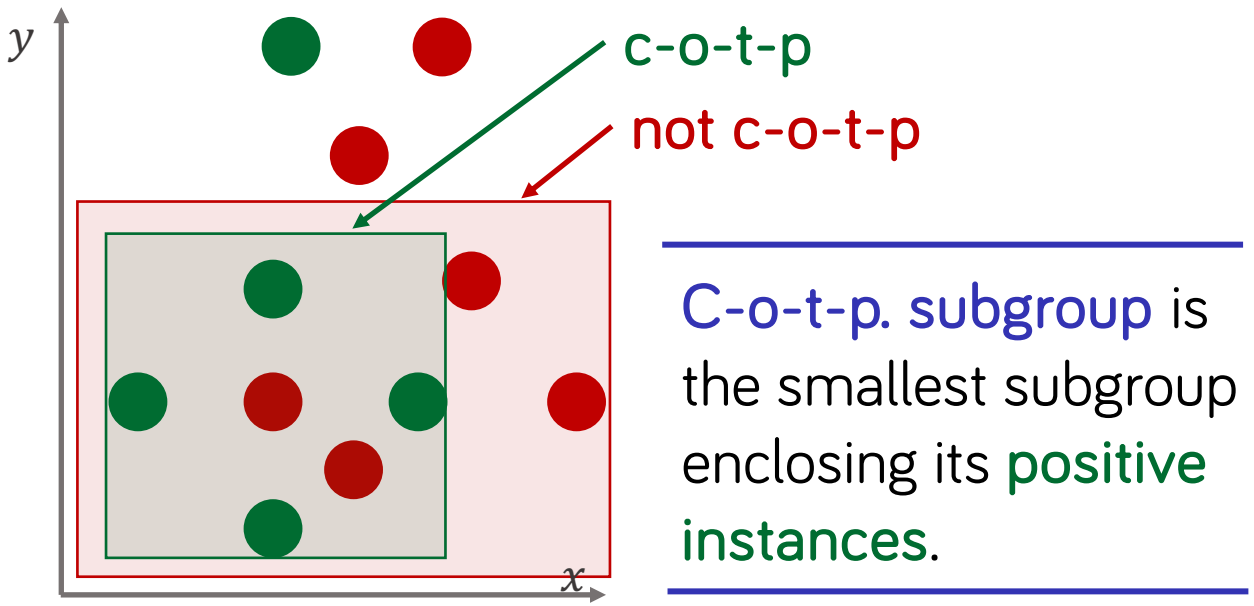
Add 3 Optimizations

H. Grosskreutz, S. Rüping, S. Wrobel. Tight Optimistic Estimates for Fast Subgroup Discovery. ECML/PKDD (2008)

Greedy Scheme – Algorithm FSSD

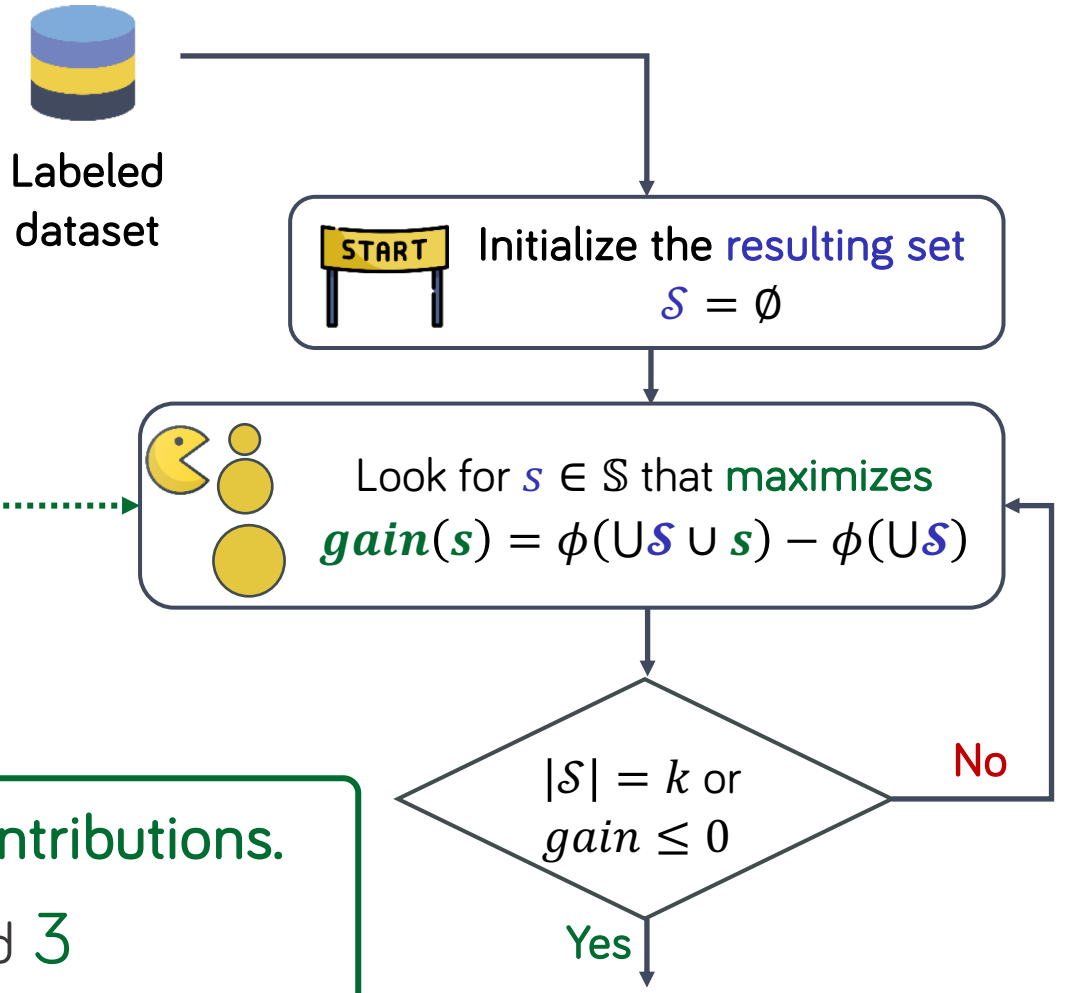
2. Ignore non closed-on-the-positives:

C-o-t-p. subgroups dominate the others.



C-o-t-p. subgroup is the smallest subgroup enclosing its positive instances.

G. C. Garriga, P. Kralj Novak, N. Lavrac. Closed Sets for Labeled Data. PKDD (2006), JMLR(2008)



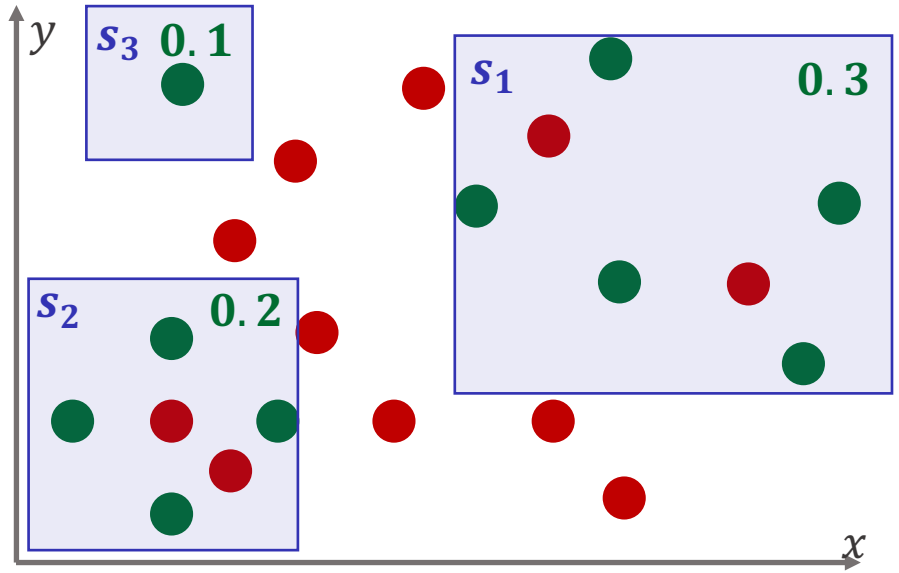
Contributions.

Add 3 Optimizations

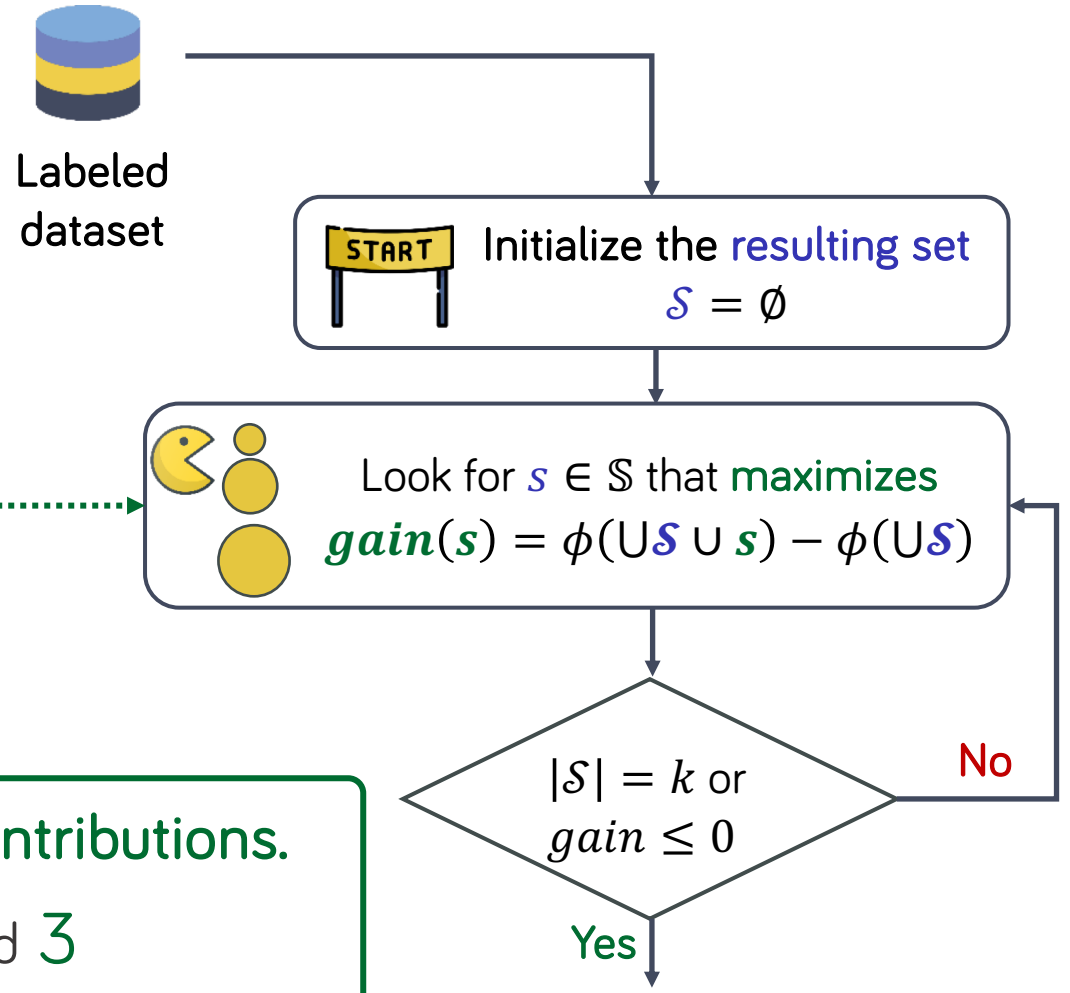
Greedy Scheme – Algorithm FSSD

3. Ignore already covered instances:

Deleting covered instances after each step DOES NOT CHANGE the result.



N. Lavrac, B. Kavsek, P. A. Flach, L. Todorovski. Subgroup Discovery with CN2-SD. JMLR (2004)



Contributions.

Add 3 Optimizations

Greedy Scheme – Algorithm FSSD

1. Ignore unpromising branches:

Tight optimistic estimates on the $gain$

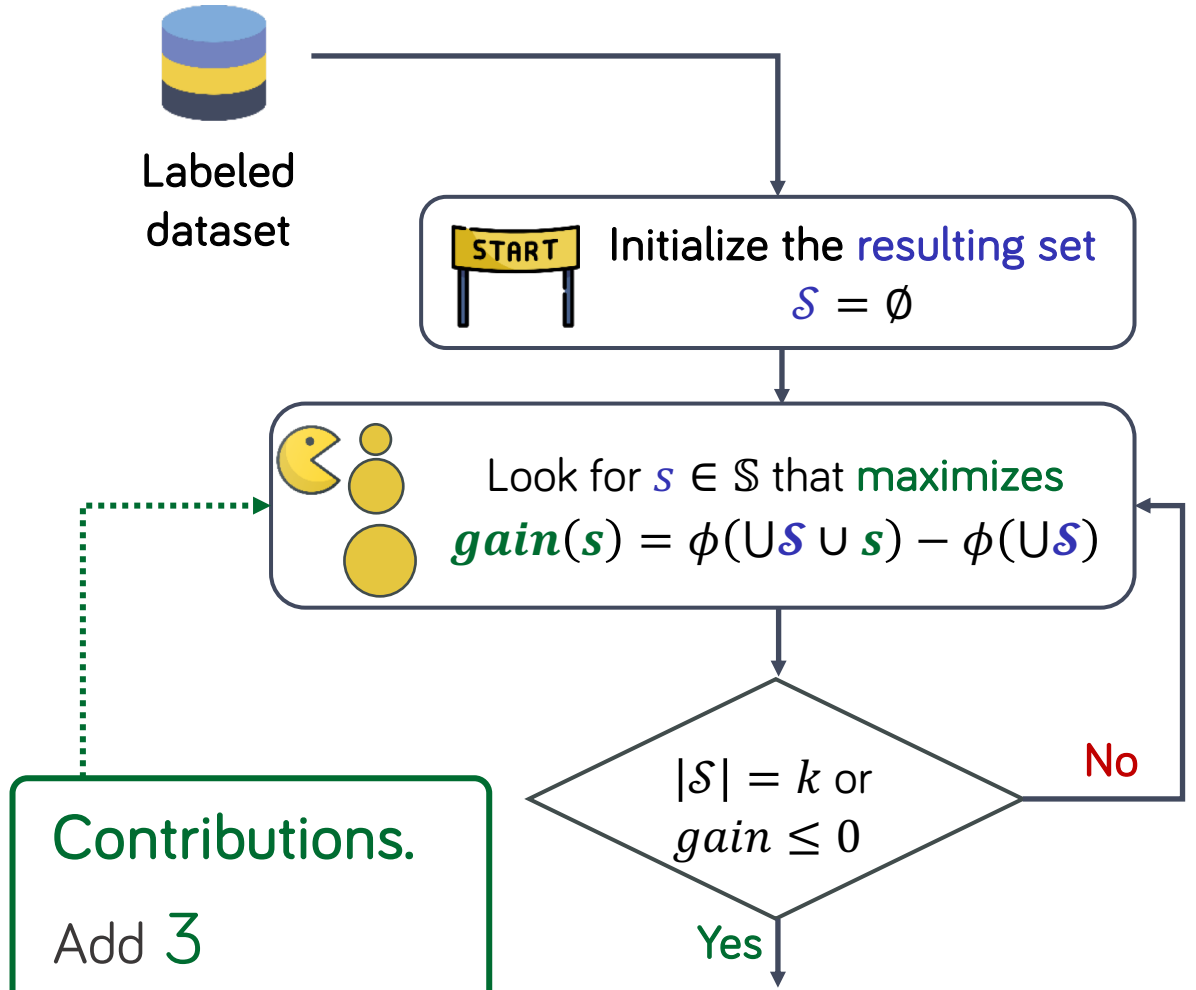
2. Ignore non closed-on-the-positives:

C-o-t-p. subgroups dominate the others.

3. Ignore already covered instances:

Deleting covered instances after each step

DOES NOT CHANGE the result.



Contributions.

Add 3

Optimizations

Outline

Greedy scheme and guarantees

For any ϵ, one can create an instance s.t.

$$\frac{\phi(\textbf{\textit{FSSD output}})}{\phi(\textbf{\textit{OPTIMAL output}})} < \epsilon$$

Approximation ratio comparing to the optimal solution

$$\rho = \frac{\phi(\textbf{\textit{FSSD output}})}{\phi(\textbf{\textit{OPTIMAL output}})}$$

$S \mapsto \phi(US)$ is NOT a submodular set function in the general case.

(e.g. proof is trivial for ϕ Informedness, WRAcc).

Do we have guarantees on the approximation ratio for FSSD? No

No guarantees for FSSD ...

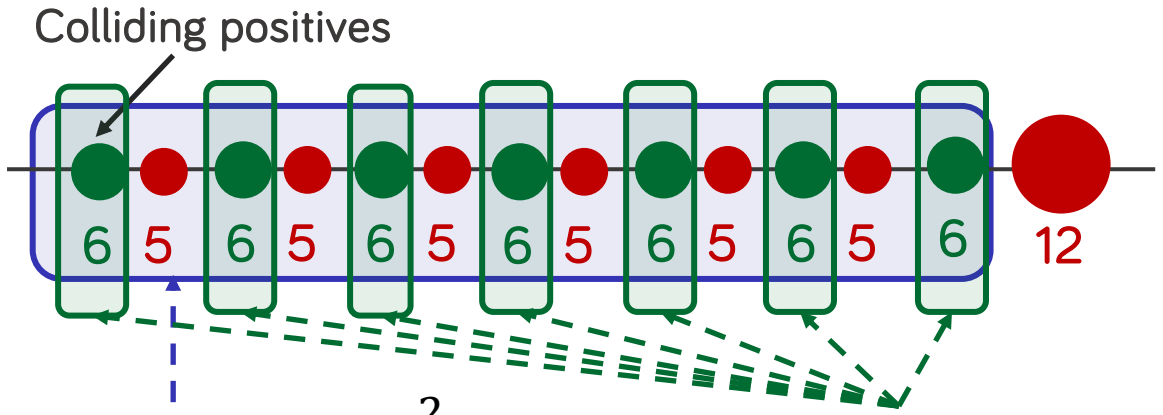
For any ϵ, one can create an instance s.t.

$$\frac{\phi(\textit{FSSD output})}{\phi(\textit{OPTIMAL output})} < \epsilon$$

Example. Fix ϵ to 30%

$\phi = tpr - fpr$ and $k = 7$

Subgroups are all subsets selected by an interval



$\phi(\textit{FSSD output}) = \frac{2}{7} \approx 28.5\%$ $\phi(\textit{BEST output}) = 1$

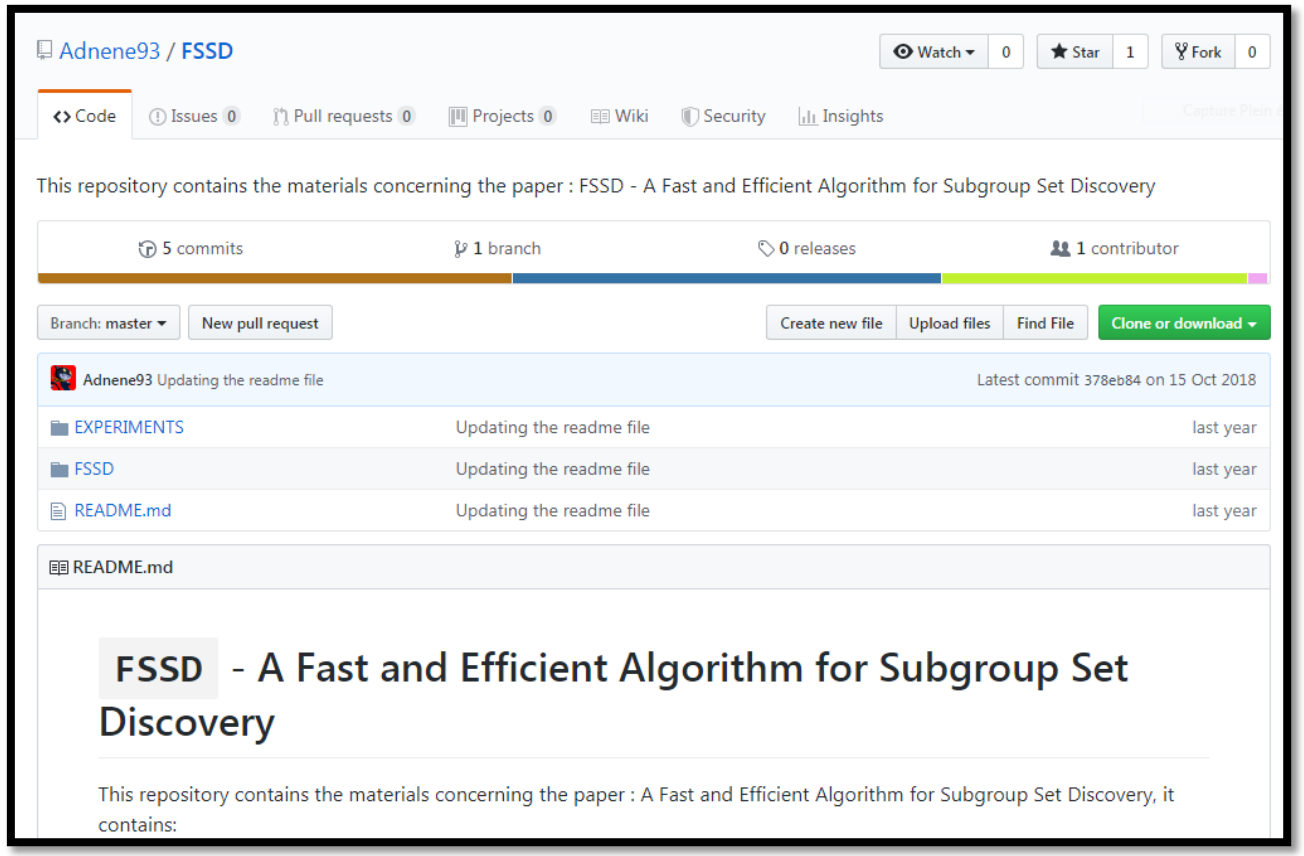
$S \mapsto \phi(US)$ is NOT a submodular set function in the general case.

(e.g. proof is trivial for ϕ Informedness, WRAcc).

Do we have guarantees on the approximation ratio for FSSD? No

Outline

Empirical Study



id	dataset	rows	class	$\frac{\lvert\mathcal{G}^+\rvert}{\lvert\mathcal{G}\rvert}$	#attrs (cat./num.)
D01	*abalone*	4177	M	0.37	(0/8)
D02	*adult*	32561	$\geq 50K$	0.24	(8/6)
D03	*autos*	195	3	0.12	(11/14)
D04	*balance*	625	B	0.08	(0/4)
D05	*breastCancer*	683	4	0.35	(0/9)
D06	*BreastTissue*	106	car	0.20	(0/10)
D07	*CMC*	1473	2	0.23	(0/9)
D08	*credit*	666	+	0.45	(9/6)
D09	*dermatology*	358	3	0.20	(0/34)
D10	*glass*	214	3	0.08	(0/10)
D11	*haberman*	306	2	0.26	(0/3)
D12	*iris*	150	V	0.33	(0/4)
D13	*mushrooms*	8124	p	0.48	(22/0)
D14	*sonar*	208	R	0.47	(0/60)
D15	*TicTacToe*	958	-	0.35	(9/0)

Source code in Python 3.7.2

https://github.com/Adnene93/FSSD

Conducted an empirical study on

15 UCI Datasets

Q1 – How good are the results provided by FSSD?

id	ρ	id	ρ	id	ρ	id	ρ
D01-1	99.7%	D05-1	100%	D09-5	99.7%	D13-7	99.6%
D02-3	100 %	D06-1	90.54%	D10-3	62.2%	D14-1	100 %
D03-7	87.3%	D07-4	99.66%	D11-1	100%	D15-4	100 %
D04-2	100 %	D08-7	99.62%	D12-1	100 %		

Approximation ratio comparing to the optimal solution

$$\rho = \frac{\phi(\boldsymbol{FSSD\ output})}{\phi(\boldsymbol{OPTIMAL\ output})}$$

In 14 out of 15 datasets:

$$\rho \geq 87\%$$

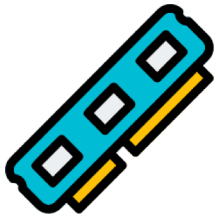
Q2 – How fast and efficient is FSSD?

id	BASELINE		FSSD	
	t(s)	M(MiB)	t(s)	Mem.(MiB)
D01-5	0.37	22.73	**0.21**	**15.34**
D02-10	120.31	426.16	**8.70**	**40.09**
D03-10	34.37	37.10	**0.09**	**9.13**
D04-4	0.77	25.60	**0.29**	**1.55**
D05-9	2.63	103.52	**0.06**	**2.58**
D06-10	38.01	23.98	**0.04**	**5.51**
D07-9	97.54	2556.15	**45.62**	**7.42**
D08-10	3.76	77.22	**0.26**	**3.33**
D09-10	34.46	338.61	**0.06**	**2.32**
D10-10	11.89	26.11	**0.07**	**9.46**
D11-3	33.78	1141.01	**7.16**	**2.58**
D12-4	15.78	341.79	**0.02**	**1.29**
D13-10	149.25	457.50	**3.96**	**33.52**
D14-10	36.97	1533.13	**17.77**	**4.08**
D15-9	2.68	20.77	**0.27**	**2.06**

BASELINE: Naïve greedy scheme

1. Extract ALL subgroups then
2. Build the subgroup set greedily

FSSD is ~200 times faster than BASELINE in average.



FSSD consumes ~10 times less memory than BASELINE in average.

Q3 – How FSSD behaves comparing to the other algorithms?

id	DSSD		MCTS4DM		CN2SD		FSSD	
	t(s)	Qual	t(s)	Qual	t(s)	Qual	t(s)	Qual
D01-5	**4.93**	0.06	30.28	0.05	11.75	0.04	72.09	**0.07**
D02-10	504	0.10	111.53	-	130	0.07	**237**	**0.11**
D03-10	1.77	**0.08**	788.79	-	3.54	-	**0.01**	0.07
D04-4	1.36	0.02	3.91	0.002	7.62	-	**0.28**	**0.03**
D05-9	2.30	0.17	2.28	0.05	3.58	-	**1.18**	**0.22**
D06-10	1.83	0.14	4.40	-	1.64	-	**0.01**	**0.16**
D07-9	2.95	0.06	**2.87**	0.06	20.19	0.03	133	**0.08**
D08-10	2.34	0.18	1864	-	7.24	-	**0.23**	**0.19**
D09-10	1.40	0.09	1.76	0.004	2.58	-	**0.03**	**0.16**
D10-10	2.08	0.06	2.66	0.02	2.92	-	**0.01**	**0.07**
D11-3	1.38	0.07	5.16	0.02	4.61	0.08	**0.35**	**0.09**
D12-4	1.34	0.20	2.59	0.20	1.35	0.20	**0.01**	**0.22**
D13-10	6.56	0.19	565.81	-	2.94	-	**0.54**	**0.23**
D14-10	**2.32**	0.11	9.09	0.08	10.71	-	933	**0.16**
D15-9	1.91	0.07	178.04	-	3.24	0.13	**0.24**	**0.17**

FSSD is **faster** than its contenders in 12 out of 15 datasets.

FSSD provides the **best** solution (w.r.t. our measure) in 14 out of 15 datasets.

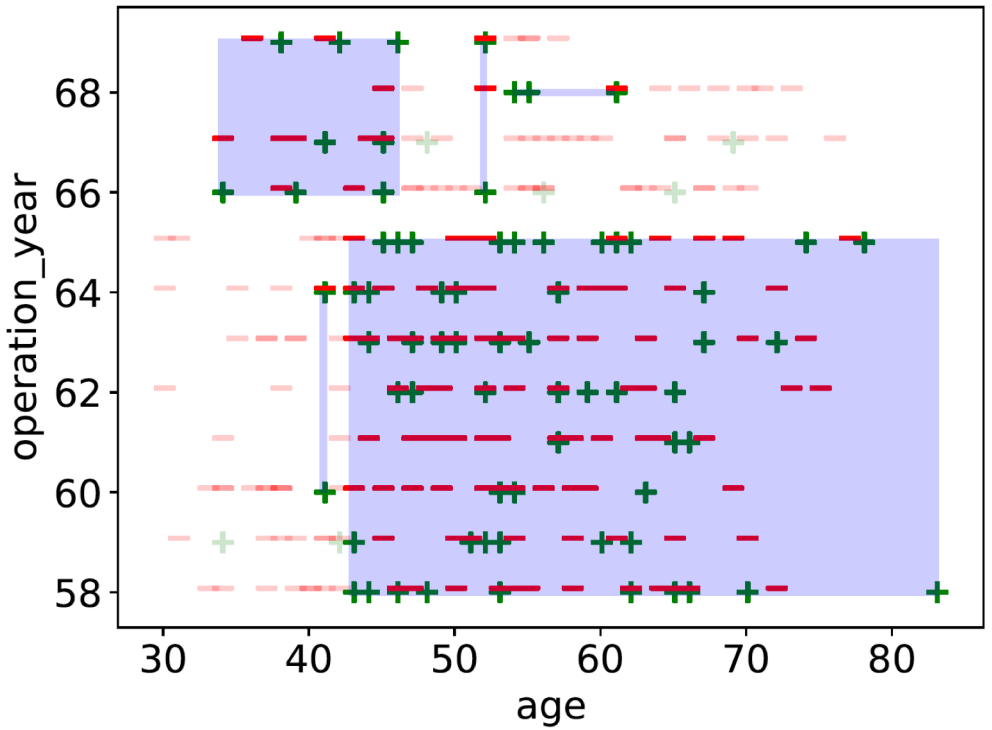
N. Lavrac, B. Kavsek, P. A. Flach, L. Todorovski. Subgroup Discovery with CN2-SD. JMLR 2004

M. van Leeuwen, A. Knobbe. Diverse Subgroup Set Discovery. DAMI 2012

G. Bosc, J. F. Boulicault, C. Raïssi, M. Kaytoue. Anytime discovery of a diverse set of patterns with Monte Carlo tree search. DAMI 2018

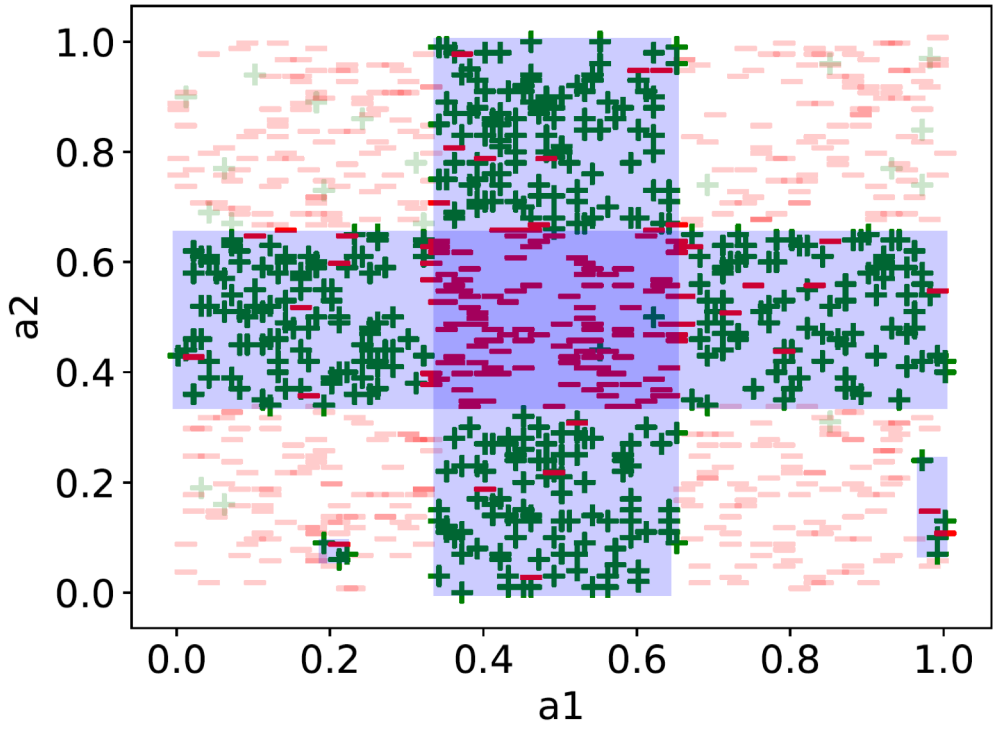
Q4 – How are the results provided by FSSD?

Haberman's Survival dataset*



Non-Overlapping subgroups

Synthetic dataset



Overlapping subgroups

*https://archive.ics.uci.edu/ml/datasets/Haberman's+Survival

Outline

Conclusion and Perspectives

Find a **Diverse** set of at most k interesting subgroups in a labeled dataset.

Avoid using an **additional redundancy measure** and **maximize**

$$\phi\,(\textit{subgroup set } \mathcal{S}) = \phi\left(\bigcup_{s\in\mathcal{S}} s\right)$$

This optimization problem is NP-hard!

Propose FSSD that maximizes this measure **greedily**.

1. Tight optimistic estimates,
2. Sequential covering,
3. c-o-t-p. subgroups

Sadly, no approximation guarantees are provided.

Still, FSSD is Fast, efficient and provides **good results** in **practice**!

Perspectives

Algorithm: replace **c-o-t-p. subgroups** by (candidate) **relevant ones**.

Theory: Is it **tractable** to provide approximation guarantees?

H. Grosskreutz. Class Relevant Pattern Mining in Output-Polynomial Time. SDM (2012)

FSSD – A Fast and Efficient Algorithm for Subgroup Set Discovery

Authors: Adnene Belfodil and Aimene Belfodil, **Anes Bendimerad**, Philippe Lamarre, Céline Robardet, Mehdi Kaytoue, Marc Plantevit

Thank you for your attention. Questions ?

Several Icons are made by authors from www.flaticon.com