Probabilistic Rule Lists using the MDL Principle

#DS2018

John Aoga, Tias Guns, Siegfried Nijssen & Pierre Schaus

2

Motivation

Year	Month	Day	Part of Day	Minute	Door Opended
2018	October	Monday	Morning	10	■ Closed
	October	Monday			□ Opened
	October	Tuesday		30	■
	October	Tuesday		40	■
	October	Wednesday		50	□
	October	Wednesday		10	□
	October	Thursday		20	■
	October	Thursday		30	■
	October	Friday		40	■
	November	Monday		50	□
	November	Monday		10	■
	November	Tuesday	Afternoon	20	■
	November	Tuesday	Morning	30	■
	November	Wednesday	Afternoon	40	■
	November	Wednesday	Morning	50	■
	November	Thursday	Afternoon	10	■
	November	Thursday	Morning	20	□
	November	Friday	Afternoon	30	■
	December	Monday	Morning	40	■
	December	Monday	Afternoon	50	■
	December	Tuesday	Morning	10	■
	December	Tuesday	Afternoon	20	□
	December	Wednesday	Morning	30	■
	December	Wednesday	Afternoon	40	■
	December	Thursday	Morning	50	■
	December	Thursday	Afternoon	10	■
	December	Friday	Morning	20	■

Can you summarize my life based on this data in interpretable way?

3

State of the Art

# always predict default label
# specific rules for exception

Rule-based classification

- set of rules that **predicts class of examples well**
- **CN2, RIPPER, AQ, C4.5,SBRL**

# Capture only the local behaviour in the data

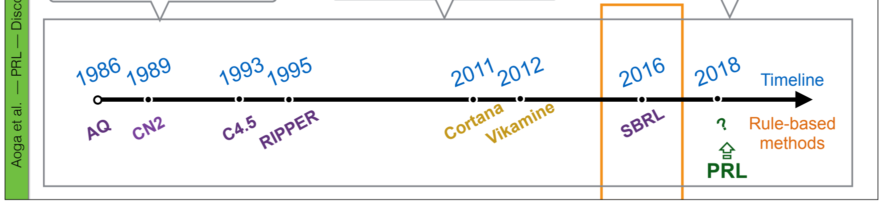
Subgroup Discovery

- set of rule that **describes subgroup of examples well**
- **Cortana, Vikamine**

# Capture only the local behaviour in the data

Probabilistic Rule List

- set of rule that **describes all examples well, being small**
- **PRL**



4

Simple Example

	🍎	🥝	🍐	🍌	♀♂
	a	b	c	e	M/F
1	✓	✓	✓	✓	M
2	✓	✓	✓	✓	F
3	✓	✓		✓	M
4		✓	✓	✓	F
5	✓		✓		F

How is the purchase of fruits depended on who buys them?

Summarize the target attribute based on what people buy in interpretable way

5 Problem of PRL

- **Given** A database of instances (observations), with for each a *Boolean target attribute*
- **Find** A Probabilistic Rule List
- **Such that** this Rule List when applied to the given database *describe* it well being *small* and *interpretable*

6 Contributions

CONTRIBUTION

Goal: Finding Rule List which learns rules with probabilities to characterize the class distribution over the entire data and favor smaller rule lists to ease interpretation

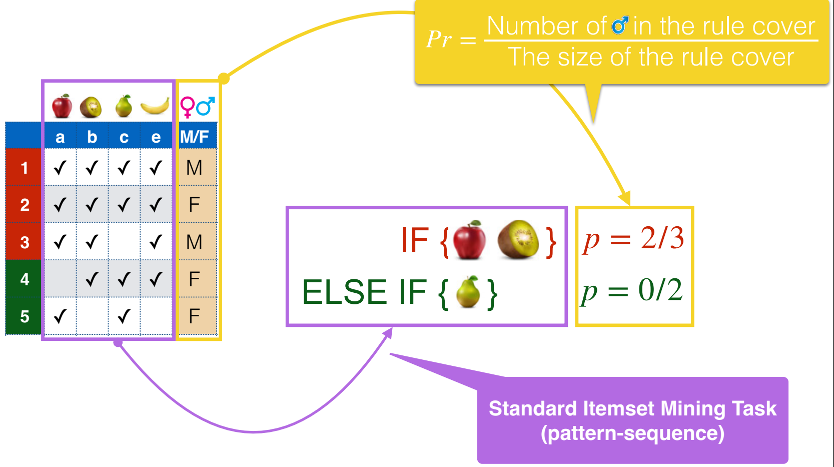
☑ New optimization criterion

- based on the MDL principle;
- aiming to find small-and-good rule lists

☑ New search algorithm

- based on branch-and-bound search;
- aiming to find the global optimum

7 Methodology (1) Pattern-Based Approach



	🍎	🥝	🍐	🍌	♀♂
	a	b	c	e	M/F
1	✓	✓	✓	✓	M
2	✓	✓	✓	✓	F
3	✓	✓		✓	M
4		✓	✓	✓	F
5	✓		✓		F

$$Pr = \frac{\text{Number of ♂ in the rule cover}}{\text{The size of the rule cover}}$$

IF {🍎 🥝} $p = 2/3$

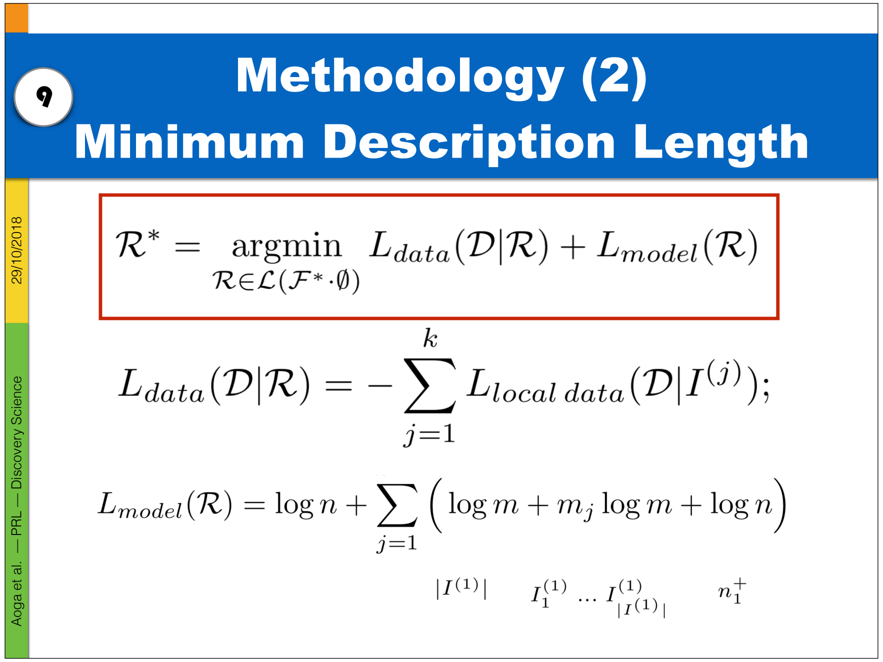
ELSE IF {🍐} $p = 0/2$

Standard Itemset Mining Task (pattern-sequence)

8 Methodology (2) Minimum Description Length

$$argmin_R \quad \text{score}(\overset{\text{Rule List}}{R}, \overset{\text{Itemset List}}{F}, \overset{\text{Database}}{D})$$

- Score is (based on Shannon's Noiseless Channel Coding Theorem)
  - the number of bit to use each rule to encode the data (using log of probabilities)
  - the number of bit to encode the rule itself



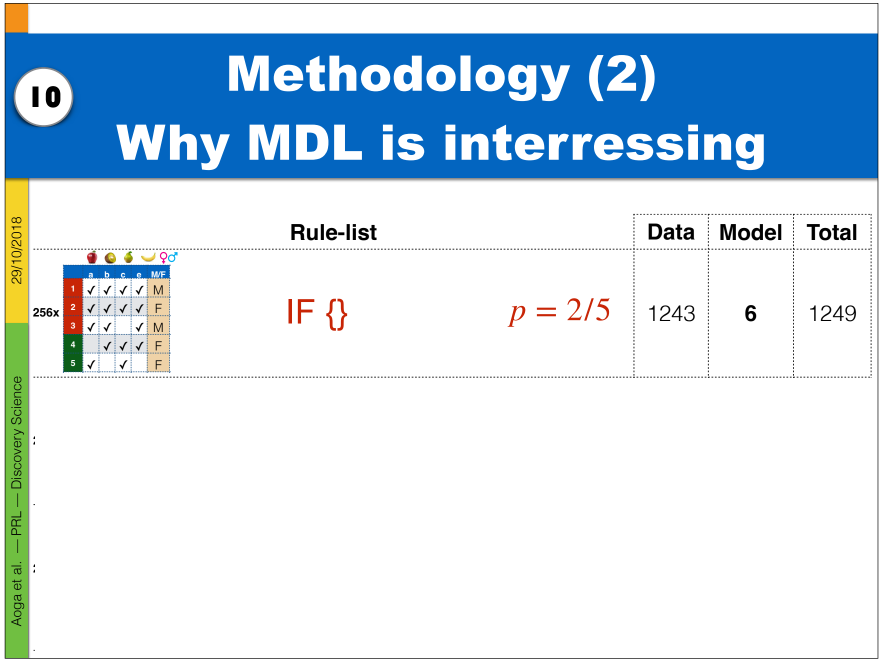
Methodology (2)
Minimum Description Length

$$\boxed{\mathcal{R}^* = \underset{\mathcal{R}\in\mathcal{L}(\mathcal{F}^*\cdot\emptyset)}{\operatorname{argmin}} \; L_{data}(\mathcal{D}|\mathcal{R}) + L_{model}(\mathcal{R})}$$

$$L_{data}(\mathcal{D}|\mathcal{R}) = -\sum_{j=1}^{k} L_{local\,data}(\mathcal{D}|I^{(j)});$$

$$L_{model}(\mathcal{R}) = \log n + \sum_{j=1} \Big(\log m + m_j \log m + \log n \Big)$$

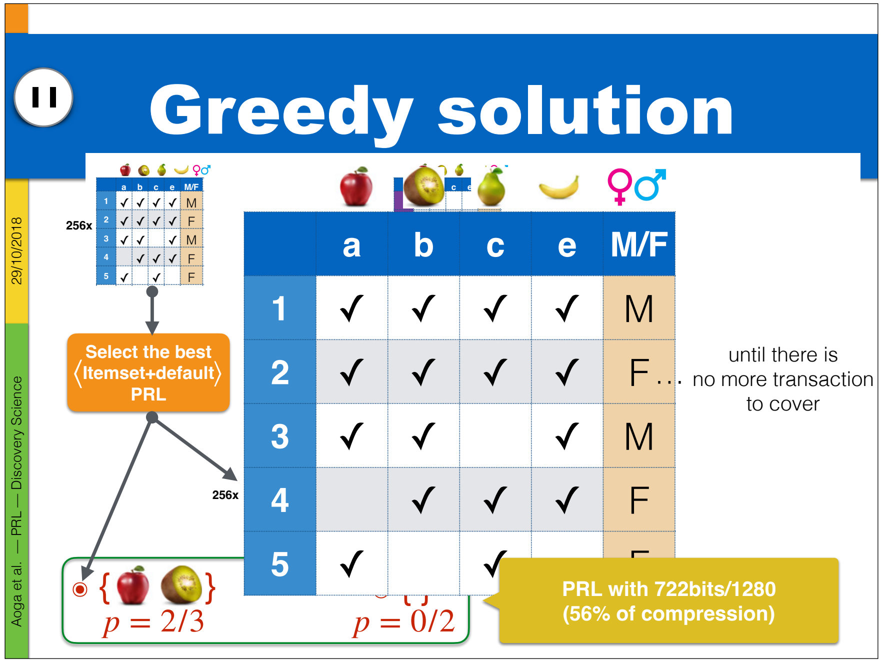
$|I^{(1)}|$ $\quad I_1^{(1)} \dots I_{|I^{(1)}|}^{(1)}$ $\quad n_1^+$



Methodology (2)
Why MDL is interressing

	Rule-list		Data	Model	Total
256x	IF {}	$p = 2/5$	1243	6	1249

	a	b	c	e	M/F
1	✓	✓	✓	✓	M
2	✓	✓	✓	✓	F
3	✓	✓		✓	M
4		✓	✓	✓	F
5	✓		✓		F



Greedy solution

256x

	a	b	c	e	M/F
1	✓	✓	✓	✓	M
2	✓	✓	✓	✓	F…
3	✓	✓		✓	M
4		✓	✓	✓	F
5	✓		✓		F

Select the best ⟨Itemset+default⟩ PRL

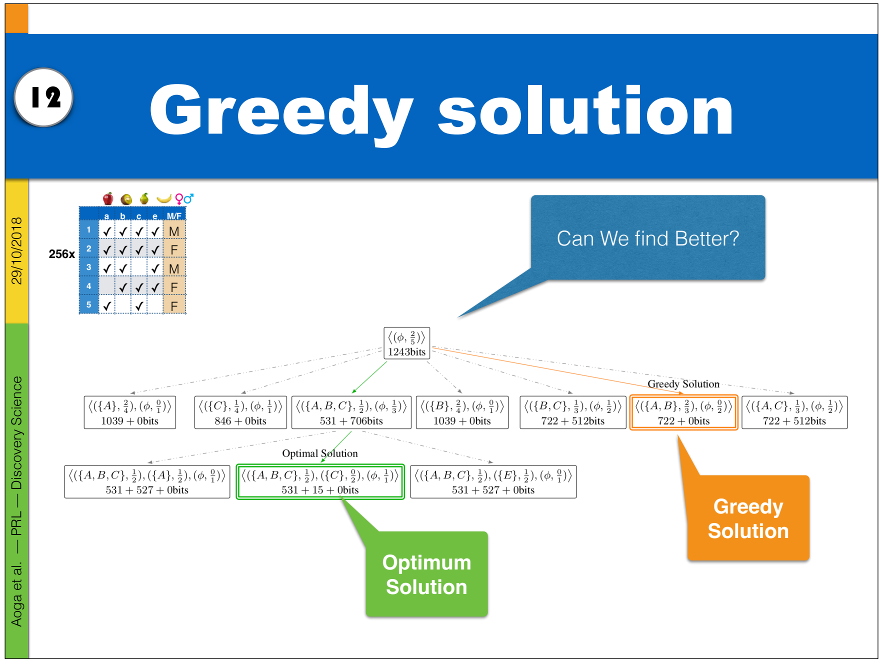
256x

until there is no more transaction to cover

{ 🍎 🥝 } $p = 2/3$

{} $p = 0/2$

PRL with 722bits/1280 (56% of compression)



Greedy solution

256x

	a	b	c	e	M/F
1	✓	✓	✓	✓	M
2	✓	✓	✓	✓	F
3	✓	✓		✓	M
4		✓	✓	✓	F
5	✓		✓		F

Can We find Better?

$\langle(\phi, \frac{2}{5})\rangle$ 1243bits

$\langle(\{A\}, \frac{2}{4}), (\phi, \frac{0}{1})\rangle$ 1039 + 0bits

$\langle(\{C\}, \frac{1}{4}), (\phi, \frac{1}{1})\rangle$ 846 + 0bits

$\langle(\{A,B,C\}, \frac{1}{2}), (\phi, \frac{1}{3})\rangle$ 531 + 706bits

$\langle(\{B\}, \frac{2}{4}), (\phi, \frac{0}{1})\rangle$ 1039 + 0bits

$\langle(\{B,C\}, \frac{1}{3}), (\phi, \frac{1}{2})\rangle$ 722 + 512bits

Greedy Solution

$\langle(\{A,B\}, \frac{2}{3}), (\phi, \frac{0}{2})\rangle$ 722 + 0bits

$\langle(\{A,C\}, \frac{1}{3}), (\phi, \frac{1}{2})\rangle$ 722 + 512bits

$\langle(\{A,B,C\}, \frac{1}{2}), (\{A\}, \frac{1}{2}), (\phi, \frac{0}{1})\rangle$ 531 + 527 + 0bits

Optimal Solution

$\langle(\{A,B,C\}, \frac{1}{2}), (\{C\}, \frac{0}{2}), (\phi, \frac{1}{1})\rangle$ 531 + 15 + 0bits

$\langle(\{A,B,C\}, \frac{1}{2}), (\{E\}, \frac{1}{2}), (\phi, \frac{0}{1})\rangle$ 531 + 527 + 0bits

Optimum Solution

Greedy Solution

Branch-and-Bound

Algorithm 2: *Branch-and-bound* $(\mathcal{F}, \mathcal{D})$

```
$PQ$ : PriorityQueue    ▷ Partial rule lists ordered by code-length when adding default rule
$best\mathcal{R} \leftarrow \langle\emptyset\rangle$, $best \leftarrow L(best\mathcal{R})$
$PQ$.enqueue-with-priority($\langle\rangle$, $L(\langle\emptyset\rangle)$)
while $\mathcal{R} \leftarrow PQ.dequeue()$ do
    for each $I \in \mathcal{F} \setminus \mathcal{R}$ do
        $\mathcal{R}' \leftarrow \langle \mathcal{R}, I \rangle$
        if $L(\langle \mathcal{R}', \emptyset \rangle) < best$ then
            $best\mathcal{R} = \langle \mathcal{R}', \emptyset \rangle$, $best \leftarrow L(best\mathcal{R})$
        if $lower\text{-}bound(\mathcal{R}') < best$ then
            $PQ$.enqueue-with-priority($\mathcal{R}'$, $L(\langle \mathcal{R}', \emptyset \rangle)$)
return $best\mathcal{R}$
```

Start by default rule list

Add iteratively new rule in the rule List

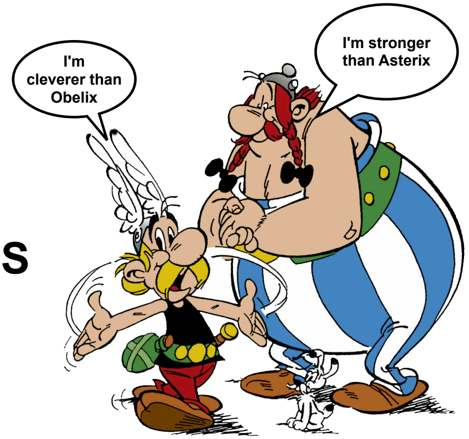
update the best if the new rule + default has minimum length than the curent best

Compute the bound and store expandable rule-list in the PQ

Lower bound computation

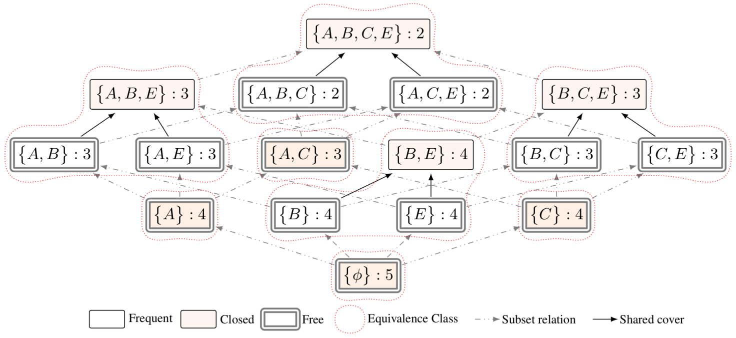
- **When we have a partial rule list, can we remove some possibilities?**
- A good lower-bound is difficult to compute since there is an exponential number of rules that can be added to the list
- In the perfect case,
  - any expansion has to be greater than or equal in size to 1,
  - and any expansion will achieve at best a data compression of 0

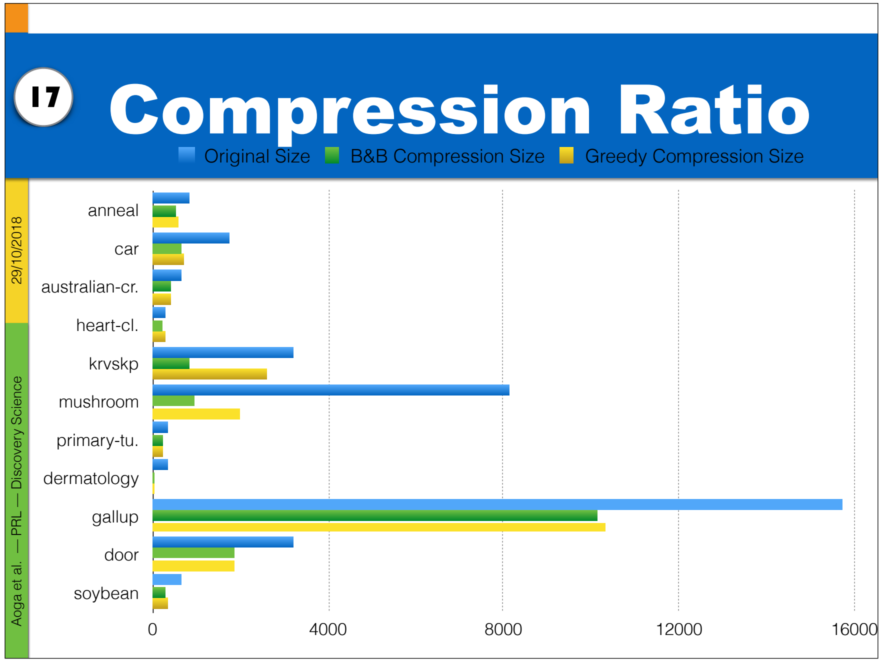
Experiments



Implementation Details

- Set representation as a Bitvector + Bitwise operation
- Only find Rule in Free-sets (equivalent classes)

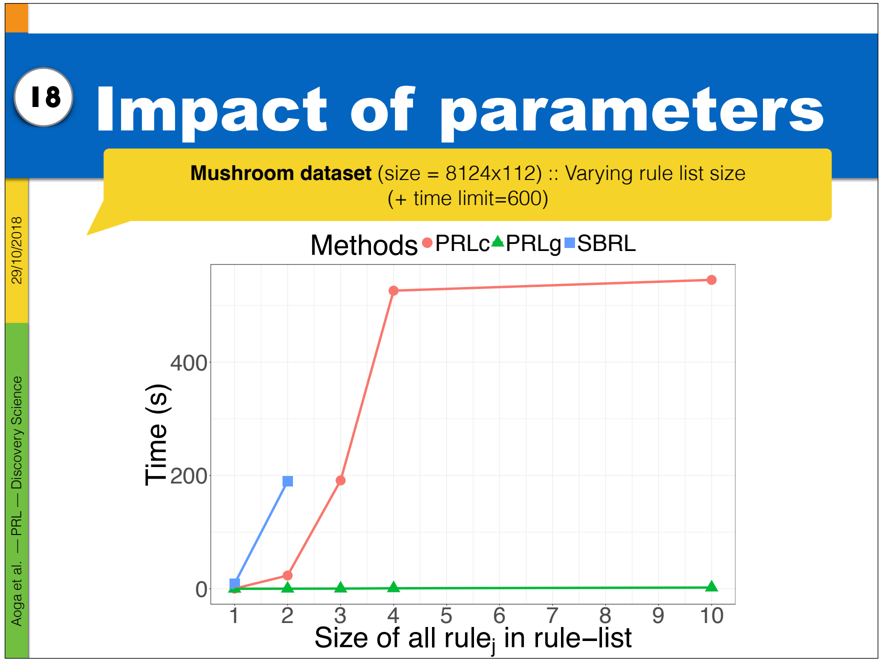




17 Compression Ratio

29/10/2018

Aoga et al. — PRL — Discovery Science

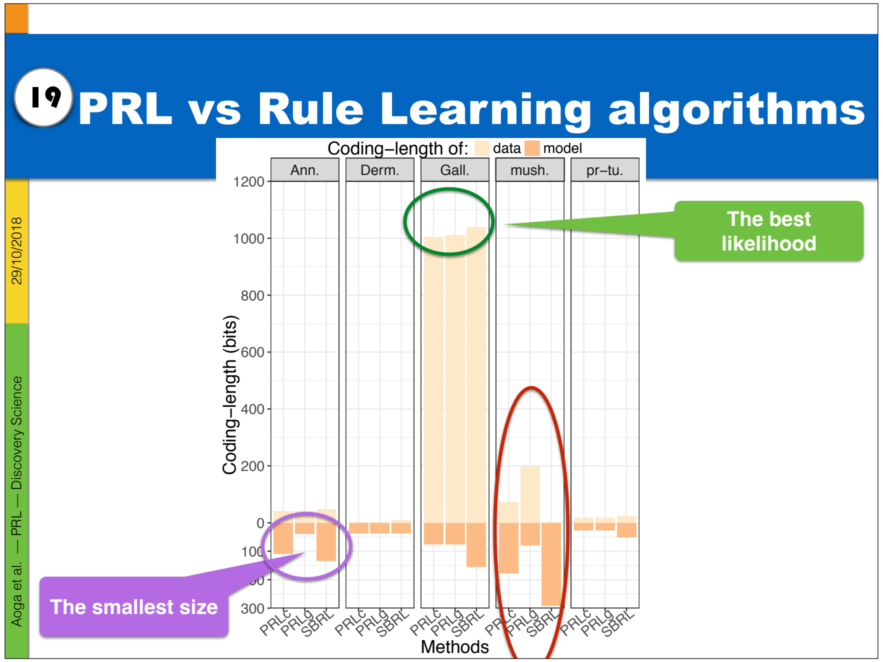


18 Impact of parameters

Mushroom dataset (size = 8124x112) :: Varying rule list size (+ time limit=600)

29/10/2018

Aoga et al. — PRL — Discovery Science



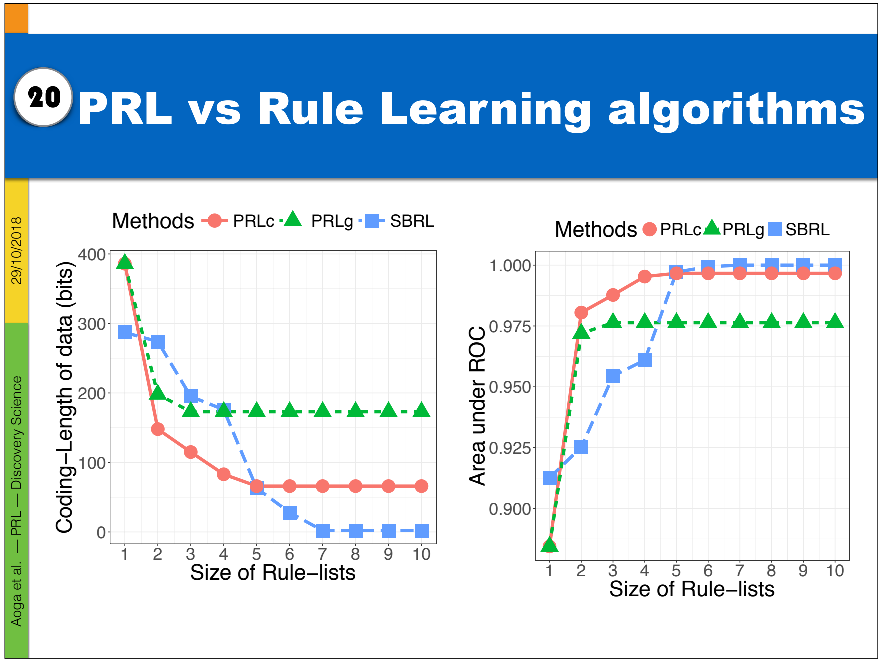
19 PRL vs Rule Learning algorithms

The best likelihood

The smallest size

29/10/2018

Aoga et al. — PRL — Discovery Science

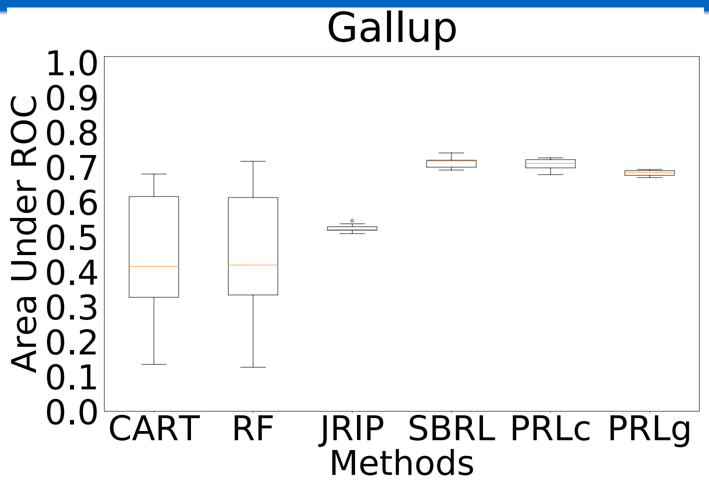


20 PRL vs Rule Learning algorithms

29/10/2018

Aoga et al. — PRL — Discovery Science

21 Prediction power of PRL



22 Conclusion

- We propose a New Descriptive method called Probabilistic Rule List
  - This Rule List is designed to be small and characterize well the target data
  - We found using a new optimization criterion based on MDL principle
  - We also designed a branch-and-bound method using Best First Search strategy

?